Teza de abilitare: AFFECTIVE COMPUTING APPLIED TO VIRTUAL REALITY BASED-PHOBIA TREATMENT AND LEARNING PROCESS

Autor: Conf. dr. Gabriela Moise

Domeniul de doctorat: Informatică

Portofoliul de lucrări relevante din domeniul de doctorat Informatică

1. Mitruț, O., **Moise, G.**, Moldoveanu, A. et al. Clarity in complexity: how aggregating explanations resolves the disagreement problem. Artif Intell Rev 57, 338, 2024. https://doi.org/10.1007/s10462-024-10952-7

   https://link.springer.com/article/10.1007/s10462-024-10952-7

2. Balan, O., **Moise, G.**, Moldoveanu, A., Moldoveanu, F. and Leordeanu, M., Automatic Adaptation of Exposure Intensity in VR Acrophobia Therapy, Based on Deep Neural Networks. In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 Research Papers. https://aisel.aisnet.org/ecis2019_rp/52
3. Bălan, O.; **Moise, G.**; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. An Investigation of Various Machine and Deep Learning Techniques Applied in Automatic Fear Level Detection and Acrophobia Virtual Therapy. Sensors 2020, 20, 496. https://doi.org/10.3390/s20020496. https://www.mdpi.com/1424-8220/20/2/496
4. Bălan, O., Cristea, Ș., Moldoveanu, A., **Moise, G.**, Leordeanu, M., Moldoveanu, F. (2020). Towards a Human-Centered Approach for VRET Systems: Case Study for Acrophobia. In: Siarheyeva, A., Barry, C., Lang, M., Linger, H., Schneider, C. (eds) Advances in Information Systems Development. ISD 2019. Lecture Notes in Information Systems and Organisation, vol 39. Springer, Cham. https://doi.org/10.1007/978-3-030-49644-9_11 https://aisel.aisnet.org/isd2014/proceedings2019/NewMedia/ https://link.springer.com/chapter/10.1007/978-3-030-49644-9_11
5. **Moise, G.**, Nicoară, E., S., Chapter 4 - Ethical aspects of automatic emotion recognition in online learning, Editor(s): Santi Caballé, Joan Casas-Roma, Jordi Conesa, In Intelligent Data-Centric Systems, Ethics in Online AI-based Systems, Academic Press, 2024, Pages 71-95, ISBN 9780443188510, https://doi.org/10.1016/B978-0-443-18851-0.00003-2 https://www.sciencedirect.com/science/article/abs/pii/B9780443188510000032?via%3Dihub

6. **Moise, G.**, Dragomir, E.G., Șchiopu, D. et al. Towards Integrating Automatic Emotion Recognition in Education: A Deep Learning Model Based on 5 EEG Channels. Int J Comput Intell Syst 17, 230, 2024. https://doi.org/10.1007/s44196-024-00638-x https://link.springer.com/article/10.1007/s44196-024-00638-x
7. Vladoiu, M., **Moise, G.**, & Constantinescu, Z., Towards Building Creative Collaborative Learning Groups Using Reinforcement Learning. In B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Designing Digitalization (ISD2018 Proceedings). Lund, Sweden: Lund University. ISBN: 978-91-7753-876-9. 2018. http://aisel.aisnet.org/isd2014/proceedings2018/Education/9
8. **Moise, G.**, Vladoiu, M., & Constantinescu, Z. (2018). Towards Construction of Creative Collaborative Teams Using Multiagent Systems. In B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Designing Digitalization (ISD2018 Proceedings). Lund, Sweden: Lund University. ISBN: 978-91-7753-876-9. 2018. http://aisel.aisnet.org/isd2014/proceedings2018/Education/10
9. **Moise, G.**, Vladoiu, M., Constantinescu, Z., Building the Most Creative and Innovative Collaborative Groups Using Bayes Classifiers. In: Panetto, H., et al. On the Move to Meaningful Internet Systems. OTM 2017 Conferences. International Conference on Cooperative Information Systems (CoopIS) 2017, Lecture Notes in Computer Science(), vol 10573. Springer, Cham. https://doi.org/10.1007/978-3-319-69462-7_17 https://link.springer.com/chapter/10.1007/978-3-319-69462-7_17

Data: 5 mai 2025

Artificial Intelligence Review (2024) 57:338
https://doi.org/10.1007/s10462-024-10952-7

Clarity in complexity: how aggregating explanations resolves the disagreement problem

Oana Mitruț[1] · Gabriela Moise[2] · Alin Moldoveanu[1] · Florica Moldoveanu[1] · Marius Leordeanu[1] · Livia Petrescu[3]

Accepted: 10 September 2024 / Published online: 19 October 2024

Abstract
The Rashômon Effect, applied in Explainable Machine Learning, refers to the disagreement between the explanations provided by various attribution explainers and to the dissimilarity across multiple explanations generated by a particular explainer for a single instance from the dataset (differences between feature importances and their associated signs and ranks), an undesirable outcome especially in sensitive domains such as healthcare or finance. We propose a method inspired from textual-case based reasoning for aligning explanations from various explainers in order to resolve the disagreement and dissimilarity problems. We iteratively generated a number of 100 explanations for each instance from six popular datasets, using three prevalent feature attribution explainers: LIME, Anchors and SHAP (with the variations Tree SHAP and Kernel SHAP) and consequently applied a global cluster-based aggregation strategy that quantifies alignment and reveals similarities and associations between explanations. We evaluated our method by weighting the k-NN algorithm with agreed feature overlap explanation weights and compared it to a non-weighted k-NN predictor, having as task binary classification. Also, we compared the results of the weighted k-NN algorithm using aggregated feature overlap explanation weights to the weighted k-NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors). Our global alignment method benefited the most from a hybridization with feature importance scores (information gain), that was essential for acquiring a more accurate estimate of disagreement, for enabling explainers to reach a consensus across multiple explanations and for supporting effective model learning through improved classification performance.

Keywords Explainability · Classification · Case based reasoning · Disagreement problem · Rashômon effect

Extended author information available on the last page of the article

1 Introduction

Artificial intelligence (AI) has made its way through all domains of activity offering at the same time benefits and concerns both at individual and society levels. Many studies related to AI impact assessment have been conducted to identify benefits and disadvantages of employing AI in people's lives, economy, health, politics, education, entertainment or science (Stahl et al. 2023; Malinka et al. 2023; Wolff et al. 2020; Mikalef and Gupta 2021). Explainable AI (XAI) has a major role in mitigating the risks of AI systems by increasing their transparency and trust. XAI methods are needed in order to understand the predictions and decisions of AI systems and represent a prevailing topic in the field of artificial intelligence. However, there are doubts about the reliability of explanations, as they are prone to noise and their attributes can vary substantially across techniques. Moreover, multiple explanations generated by the same feature attribution explainer can be dissimilar, creating confusion and skepticism even for AI professionals. This paper presents a method inspired from textual-case based reasoning for aggregating explanations from various explainers in order to solve the disagreement between explainers and the dissimilarity across explanations, based on the "case base image" metaphor that compares problem and solution space clusters. Thus, we seek a modality to enable the explainers to reach a consensus by exploring the depths of their feature attribution explanations and to unearth common patterns, regularities and associations (Raghunandan et al. 2008), accentuating the agreement and hiding the disagreements.

We evaluated our method by weighting the k -NN algorithm with agreed feature overlap explanation weights and compared it to a non-weighted k -NN predictor, having as task binary classification for 6 popular datasets. Also, we compared the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors).

An interesting contribution to disagreement measurement, consensus settlement and classification performance was achieved by synthesizing global feature attribution averages and global feature alignment weights with the information gain of each feature (feature importance). The idea of giving each feature a weight value corresponding to its information gain proved to be valid in our aggregation strategy across explainers and explanations as well. The intuition of merging feature importances with global averages or global alignments across explainers and their multiple explanations generated iteratively is more successful than local alignment or the simple mean of feature rankings proposed by previous studies (Pirie et al. 2023). This was very important for capturing a more accurate estimate of disagreement, for enabling explainers to reach a consensus across multiple explanations and for supporting effective model learning through improved classification performance.

In what concerns the comparison between the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), for each feature alignment scheme R, S and SR, we observed that the aggregation strategy helps the explainers to reach a consensus and resolve the disagreement problem more effectively than using a single explanation method.

We constructed our scientific investigation not only on our intuition, but also on results of Chen and Hao (2017), who demonstrated that features contribute differently to classifica-

tion – some are relevant, some are trivial relevant and others are irrelevant. Thus, we aim to reduce fluctuations in predictions from a set of models by minimizing the variability of feature scores.

The manuscript is structured as follows: Chap. 2 presents a classification of explanation models, Chap. 3 details the Rashômon Effect, the disagreement problem and a method for solutioning the disagreement problem using Case Based Reasoning, Chap. 4 describes the proposed method, Chap. 5 introduces the evaluation of the method and its consequent results, Chap. 6 exhaustively discusses the findings of the study and finally, Chap. 7, outlines the conclusions and future research directions.

2 Explaining machine learning models

2.1 General classification of explainable models

Machine learning (ML) models can be explained primarily using two approaches. The first approach is based on developing inherently interpretable models (Brughmans et al. 2023; Adadi and Berrada 2018). These models consist of low-complexity ML models, as decision trees, linear regression, logistic regression, generalized additive models. A major drawback of using these types of models is their low predictive power. Complex models such as deep neural networks offer high accuracy and reliable results. Thus, another approach to explain models is by analyzing their behavior after training (Brughmans et al. 2023; Krishna et al. 2023). The post-hoc explainability methods can be specific to certain machine learning models, such as neural networks and consider their internal structure, others do not consider the architecture of the ML models and can be used for a larger set of models (Poiret et al. 2023). The latter are called agnostic methods. In (Adadi and Berrada 2018), the intrinsic methods are viewed as specific-method and agnostic-methods are in generally tied by post-hoc explanations. The explanations can be global in the situation when we try to understand the reasoning for getting all possible results or local in the case we try to provide the reasoning for getting a single result or a prediction for a single instance.

The local post-hoc explanation algorithms are used for individual explanations and can be classified in perturbation and gradient based methods in (Krishna et al. 2023). The perturbation-based methods consist of perturbing the inputs of the model and observing the changes in their predictions. The new instances are used to build inherently interpretable approximations. In the case of gradient-based methods, gradients with respect to some features are built to explain the predictions.

Some of the most popular post-hoc explanations methods are: Local interpretable model-agnostic explanations - LIME (Ribeiro et al. 2016), SHapley Additive exPlanations - SHAP (Lundberg and Lee 2017), Anchors (Ribeiro et al. 2018), Gradient weighted Class Activation Mapping - GradCAM (Selvaraju et al. 2020), Integrated Gradients (Sundararajan et al. 2017), SmoothGrad (Smilkov et al. 2017) for local explainable methods and Generalized Additive Models - GAM (Hastie and Tibshirani 2017) for global explainable methods. LIME, SHAP and Anchors are perturbation-based methods. GradCAM, Integrated Gradients, SmoothGrad are gradient-based methods. GAM uses, as LIME and SHAP do, inherently interpretable models to imitate the behavior of complex ML models (Velmurugan et al. 2021; Krishna et al. 2023).

The post-hoc explanation methods can be classified in example-based techniques and feature-based techniques (Brughmans et al. 2023). In the case of example-based techniques (counterfactual explanations), it is analyzed what would happen if the input of the model were modified in a certain way. LIME and SHAP are feature-based techniques, because they compute for each feature a value representing the importance of that feature in the model's prediction (attribution score) (Müller et al. 2023; Velmurugan et al. 2021). The attribution scores depend on the model, the sample and the attribution method (Müller et al. 2023). Recent research focuses on establishing an agreement between various attribution methods (Pirie et al. 2023).

Adadi and Berrada (2018) provide a comprehensive taxonomy for XAI techniques, which is resumed in Fig. 1.

2.2 Motivation for using LIME, SHAP and anchors in our research

We chose LIME, SHAP and Anchors in our research because of their popularity and availability of resources - Python libraries, packages and tutorials, visualization tools, easy to install components and straightforward programming implementation. In addition, other works that focused on establishing an agreement between explanation methods, such as Pirie et al. (2023), used LIME and SHAP in their experiments. As their approach was to define a local alignment that computes the similarity of problems and solutions in the neighborhood of individual cases and ours is to exploit global alignment, by comparing problem and solution space clusters, we considered that by using some of the same explanation methods we could facilitate a comparison between the two research ideas.

LIME and Anchors were proposed by the same research team - LIME in 2016 (Ribeiro et al. 2016) and Anchors in 2018 (Ribeiro et al. 2018). Both use perturbation-based strategies to provide local explanations for predictions generated by a black-box model. LIME uses a surrogate model to approximate the behavior of the black-box model for the values in the vicinity of an instance.

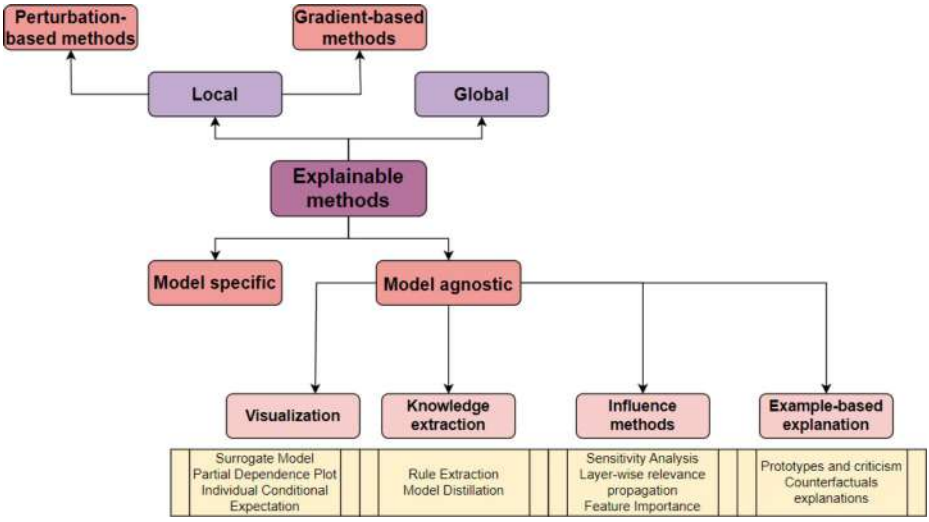

Fig. 1 General taxonomy of explanation methods

LIME can be applied to any model, does not need access to the model internals, is portable and can be used even if the original model changes, provides explanations as feature importances, uses a large number of features, its quantitative explanation visualization is more intuitive and works for tabular, text and image data (Saarela and Geogieva 2022). In comparison, although the integrated gradients method (applied by Pirie et al. 2023 for local explanations alignment) is more stable and robust, it can only be applied to differentiable models.

The SHAP method (Lundberg and Lee 2017) determines the contribution of each feature to the prediction by computing the Shapley values from the coalitional game theory. The Shapley values were introduced by Shapley (1953) to quantify the contribution of individuals to the results of a cooperative game. Similarly to Poiret et al. 2023; we chose SHAP for its ease of use, widespread popularity and axiomatic superiority. It assigns each feature an importance in a model-agnostic way and can be compared to other unstable and manipulable methods such as LIME, where each run produces a different explanation (Velmurugan et al. 2021). SHAP combines the locality of LIME with the concept of Shapley values from game theory, which decreases the computation time (Brughmans et al. 2023). Moreover, SHAP is better aligned with human intuition as measured by user studies (Lundberg and Lee 2017).

Both LIME and SHAP compute feature importances and the impact of a certain feature is measured related to the predictive outcome (Brughmans et al. 2023). A positive/negative score is attributed to each feature to underline whether the value of the features pushes the prediction towards the positive or negative class (Roy et al. 2022). We used feature importances, scores and signs as input to the feature overlap explanation schemes for explainers and explanations alignment.

Anchors or scoped rules explain the prediction as feature conditions that retain the important criteria leading to a prediction (Kundu and Hoque 2023) in the form of if-then rules, instead of a surrogate model. Its name suggests metaphorically that it "anchors" the prediction. Compared to LIME, the Anchors method provides a region of instances to describe the model's behavior (Brughmans et al. 2023).

The results of Kundu and Hoque (2023) showed that different methods generate unstable explanations for different batches of samples, leading to disagreement among explanation methods. Therefore, an agreement between explainers and explanations needs to be established.

3 Disagreement between explainers

3.1 The Rashômon effect

The "Rashômon Effect" is a concept adopted in statistics by Breiman (2001), to reflect "the multiplicity of good models". The term is borrowed from a Japanese drama film, "Rashômon", in which four people witness some crimes. Their testimonies are different, even if they present the same facts. Breiman defines the "Rashômon Effect" as the fact "that there is often a multitude of different descriptions in a class of functions giving about the same minimum error rate". Poiret et al. (2023) provide a simple definition of the Rashômon effect, applied for explainability:

Considering f_1 and f_2 two models, where f_1 and f_2 predict the same result for an input x. The Rashômon effect means that the same predictions are obtained even if the importances of the features are different.

Let be $rank_1$ and $rank_2$ the features' ranks of f_1 and f_2. The Rashômon effect occurs when $rank_1 \neq rank_2$.

The presence of the Rashômon effect in model prediction explanations leads to a loss of confidence in the results and in the models themselves. Often, different attribution scores are obtained across different attribution methods (Müller et al. 2023). The Rashômon effect can be seen as a manifestation of underspecification, where the model is unable to seize all the underlying patterns in the data accurately (Poiret et al. 2023).

3.2 The disagreement problem

The disagreement problem refers to obtaining contradictory explanations when different explanation methods to the same model and for the same sample are used (Brughmans et al. 2023; Müller et al. 2023; Krishna et al. 2023). LIME and SHAP may provide explanations that are inconsistent (even contradicting), unstable and sensitive to adversarial attacks and fairwashing (Krishna et al. 2023; Aïvodji et al. 2019; Ghorbani et al. 2019; Slack et al. 2020). This poses a challenge to data scientists who usually base their decisions on the explanations generated by multiple such methods, because of the lack of ground truth. When the explanations are consistent, ML practitioners get a coherent understanding of the algorithm's prediction, but when not, they need to cautiously align all the explanations and eventually choose a method that would minimize all the possible risks, especially when it comes to high-stakes decisions in critical domains (medicine, law, finance). Del Giudice (2021) proposes seeking consensus among models with different assumptions or biases.

Two major questions arise: (1) How big the disagreement between two explanations is and how we measure it and (2) How we solve this disagreement in order to obtain a reliable explanation.

Müller et al. (2023) used attribution scores for explaining model predictions and proposed a global averaging dissimilarity measure for the following metrics: (1) feature disagreement; (2) sign disagreement; (3) the Euclidean distance over two attribution scores and (4) the Euclidean distance over absolute attribution scores. The study was applied on four datasets and five attribution methods (Vanilla Grad, Smooth Grad, Integrated Gradients, Kernel SHAP, and LIME) and showed the inconsistency of attribution ranking among explainers.

Krishna et al. (2023) formalized the notion of disagreement between explanations and introduced a quantitative framework to empirically analyze the extent of disagreement between the explanations generated by six state-of-the-art post hoc explanation methods. Explanation disagreement was defined by the extent to which explanations differ in the top-k features, the signs and orderings of the top-k features and the relative ordering of certain features of interest. They proposed the following metrics: feature agreement, rank agreement (magnitude of feature score), sign agreement, signed rank agreement, rank correlation and pairwise rank agreement. Higher agreement between explanation methods has been observed for the feature agreement and pairwise rank agreement metrics. Moreover, as the number of top-k features increases, rank agreement and signed rank agreement decrease. Roy et al. (2022) proposed an aggregation scheme to combine LIME and SHAP explana-

tions in the context of defect prediction. They investigated feature agreement, rank agreement and sign agreement and it was proven that top-k rank agreement is weaker than sign agreement. The alignment approach suggested presenting only the top-k most important features, common to both SHAP and LIME that share the same sign score (positive/negative) – whether the value of the feature influences the prediction towards 1 or 0 in the case of binary classification. Neely et al. (2021) observed low overall agreement between explainers and concluded that rank correlation is not a reliable metric in the absence of ground-truth rankings. Camburu et al. (2019) noticed that LIME and SHAP failed to select the relevant features and Yalcin et al. (2021) showed that the performance of Tree SHAP is inversely correlated with dataset complexity when ground-truth rankings of feature importance are provided.

Poiret et al. (2023) evaluated the similarity between multiple SHAP explanations, that were part of a Rashômon set, for different ML models. As metrics, they used the top-k features and weighted cosine similarity. The findings of the study showed that sample size affects explainability: "1) larger data volumes attenuate the Rashômon effect and improve explanation consensus, 2) explanations derived from limited data may be spurious and require validation, 3) bagging ensembles can enhance agreement between models" (Poiret et al. 2023).

Disagreement between post-hoc explanations can produce serious manipulations in adversarial contexts. In (Bordt et al. 2022), the authors demonstrated the failure of post-hoc explanations algorithms to achieve the honesty required by the legislation and the moral rules of society. They investigated the explainability algorithms in two contexts. In a "cooperative context", the two parties involved in the explanation process, the provider and the recipient of the explanation, have the same goal and interest: to find the most suitable explanation algorithm for the machine learning model. In an "adversarial context", the two parties have opposing goals and interests. For example, the representative of a bank and the customer of the bank, to whom the bank refused a loan and wants to challenge the decision as being discriminatory. The provider of an explanation needs to generate "explanations that cannot be contested by the data subject or an examiner" (Bordt et al. 2022). Four key points that allow challenging the explanation's results were identified: the choice of an explanation algorithm and its particular parameters; the exact shape of the high-dimensional decision boundary; and, when applicable, the choice of the reference dataset. The analyzed scenarios proved that these algorithms could not find the "unique, true reason" of the predictions obtained by a machine learning model. Moreover, in the case of complex models, "a true reason simply does not exist".

Various strategies can be employed to manipulate the explanations, as shown in (Goethals et al. 2023): in the process stage through direct manipulations on the data and machine learning models and in the post-process stage through switching to a convenient explanation algorithm and changing the parameters of the explanation algorithm, by exploiting the non-deterministic components of the algorithms. The disagreement problem can be exploited to achieve unethical objectives of the explanation providers through fairwashing, avoiding taking responsibility for erroneous decisions, promoting computational propaganda, implementing discriminatory practices, increasing the profit through advertising and usage of certain explanations which offer the highest profit.

Fairwashing occurs when the explanations are manipulated to cover up the unfairness of the underlying machine learning models. Shamsabadi et al. (2022) investigated the fairwash-

ing theory and implemented the first method for fairwashing detection, FRAUD-Detect. The empirical results of their method demonstrated the viability and robustness of the solution.

3.3 Resolving disagreement using case based reasoning

The Case Based Reasoning (CBR) paradigm uses the experience gathered from previous problems to solve a new problem and consists of four steps: Retrieve, Reuse, Revise and Retain (4R). Bayrak and Bach (2022) applied the 4R methodology to explain the decisions of black-box models. The reason for using CBR in XAI was argued through the lens of the fact that CBR methodology meets the quality criteria required for explanations: to be trustworthy, understandable, informative, sufficient and unbiased.

Case Based Reasoning relies on the principle that "similar problems have similar solutions". The local alignment method, that computes the similarity of problems and solutions in the neighborhood of individual cases, has been approached by Pirie et al. (2023). They proposed AGREE, an explainer aggregation framework that combined the explanations of different feature attribution explainers by using the information from the neighborhood spaces of the feature attribution vectors and transforming it into explanation weights. The explanation weights have been used in a weighted k-NN algorithm whose classification accuracy (for classification datasets) / Mean Squared Error (MSE) (for regression datasets) has been compared to a non-weighted k-NN. The results showed that the case alignment confidence metric outperformed mean feature ranking in what concerns estimating the degree of disagreement and it was robust against high dimensionality.

4 Method

Subsection (4.1) presents a general description of the method, accompanied by the algorithms' pseudocodes and a comprehensive illustration of the research approach, (4.2) describes the datasets' characteristics, (4.3) details how explanations are generated, (4.4) describes how aggregation is performed by rank average, (4.5) presents the Global Alignment Measure method and subsection (4.6) illustrates how we used it in our aggregation strategy.

4.1 General description

State-of-the-art explainers such as LIME, SHAP and Anchors can produce discrepant explanations and feature attribution vectors even when the same prediction model is applied, raising doubts about the trustworthiness of the decision-making process and the practical usability of the generated explanation. Discordant explanations can occur even for a particular instance from the dataset, when the same explanation technique is applied multiple times. It has been observed that there are significant differences between the explanations generated by LIME in terms of feature rankings and signs for the same instance from the dataset, if the algorithm is run several times.

Given these considerations, we aim to resolve the disagreement problem between various explainers and the dissimilarity between multiple explanations provided by the same explainer for each instance from the dataset by developing a cluster-based aggregation

method inspired from Case-Based Reasoning. Our strategy aligns the explanations by seizing the underlying relationships between feature attribution vectors and steering them towards the point where they effectively reach a consensus.

According to CBR, similar problems have similar solutions. Local alignment, which measures the similarity of problems and solutions in the neighborhood of individual cases, has been applied for computing a metric called Case Alignment Confidence between explainers and for developing the AGREE framework (Pirie et al. 2023). On the other hand, in our aggregation strategy, we exploit global alignment, by comparing problem and solution space clusters and adopting the "case base image" metaphor that reveals patterns, regularities and associations between the case bases (Raghunandan et al. 2008).

In our algorithm, the case bases were represented by feature attribution vectors. The feature attribution vectors contained explanations (feature importance scores and signs) generated by the LIME, SHAP (Kernel SHAP and Tree SHAP) and Anchors algorithms, for the following datasets: Pima Indian Diabetes Dataset (Smith et al. 1998), Indian Liver Patient Dataset (Ramana and Venkateswarlu 2012), Hepatitis Dataset (1988), Fetal Dataset (Campos and Bernardes 2010), Abalone Dataset (Nash et al. 1994), Water Quality Dataset (Kadiwal 2021). We applied Leave-One-Out Cross-Validation (LOOCV), a procedure that estimates the performance of machine learning algorithms to make predictions on data not used to train the model (i.e., on the test set) (Brownlee 2020). So, we trained the model on the training set and obtained predictions and explanations on the test set. It requires one model to be created and evaluated for each example in the dataset. The model is trained on a training dataset consisting of $n-1$ instances from the dataset and tested on the nth instance, for which the explanation was generated. The procedure was repeated for each instance from the dataset, so for a number of n times.

For LIME, Anchors and Tree SHAP, we fitted an XGBoost binary classification model. For Kernel SHAP, we chose Support Vector Machine with the radial basis kernel.

We extracted feature attribution vectors for the 3 feature alignment schemes: feature attribution ranks (R), feature attribution signs (S) and a vector of both feature attribution ranks and feature attribution signs (SR), for each of the 6 datasets.

The algorithm is described in more detail in the following pseudocode (Fig. 2):

A measure of agreement called Global Alignment Measure estimated the alignment between problem and solution clusters of explainers and explanations. The resulting aggregated explanation weight vectors have been provided to the feature space of a weighted k -NN classifier and we compared the prediction performance against a non-weighted k -NN version. We also compared the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), for each feature alignment scheme R, S and SR. Actually, for the Anchors explanation method we had only the R feature alignment scheme, as Anchors does not compute signs in its explanations.

For evaluating and comparing the results of the feature-weighted k -NN (the features were the weights extracted from aligning the explainers and explanations) to the non-weighted k -NN algorithm, as well as for comparing the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), we split the dataset into 30% test and 70% training. The classifiers were trained on the training set and tested on the test set. We then calculated the metrics: accuracy,

```
for each explanation method E in [LIME, Anchors, TreeSHAP, KernelSHAP]:
    for each dataset D in [Diabetes, Liver, Hepatitis, Abalone, Water, Fetal]:
        for each instance i from D:
            for a number of 100 iterations:
                apply Leave-One-Out Cross-Validation

                if E == LIME or E == TreeSHAP or E == Anchors:
                    model = XGBoost classifier
                    generate prediction
                else if E == KernelSHAP:
                    model = Support Vector Machine
                    generate prediction
                if E == LIME OR E == TreeSHAP or E == KernelSHAP:
                    generate explanation
                    feature attribution vector = attribution scores + signs
                else if E == Anchors:
                    generate explanation
                    feature attribution vector = attribution scores

                calculate vector of feature attribution ranks R
                extract vector of feature attribution signs S
                extract vector of feature attribution ranks and signs SR
```

Fig. 2 Pseudocode for the algorithm that generates explanations and feature attribution vectors

F1-score and ROC-AUC score, for each of the 3 feature alignment schemes: feature attribution ranks (R), feature attribution signs (S) and a vector of both feature attribution ranks and feature attribution signs (SR), for each of the 6 datasets.

An increase from the baseline non-weighted k -NN prediction scores and single explanation k -NN scores suggest a consensus between explainers and explanations and demonstrates that feature attribution aggregation significantly improves classification performance. The algorithm is described in more detail in the following pseudocode (Fig. 3):

A comprehensive illustration of the method is presented in Fig. 4.

4.2 Datasets

We applied our aggregation method on 6 popular binary classification datasets from the domains of medicine, biology and ecology (Table 1).

The Pima Indian Diabetes Dataset (Smith et al. 1998), which contains 8 diagnostic measurements (number of pregnancies, glucose level, blood pressure, body mass index, etc.) from 768 subjects (268 with diabetes – target variable 1 and 500 without diabetes – target variable 0), predicts whether the patient suffers from diabetes or not.

The Indian Liver Patient Dataset (Ramana and Venkateswarlu 2012) contains 10 features (age, gender, total bilirubin, direct bilirubin, albumin, etc.) from 583 patients – 416 with liver disease (target variable 1) and 167 without a liver disease (target variable 0).

The Hepatitis Dataset (1988) contains 12 features (age, sex and 10 blood test results) from 615 patients – 540 blood donors (target variable 0) and 75 suffering from either hepatitis C, fibrosis or cirrhosis (target variable 1).

The Fetal Dataset (Campos and Bernardes 2010) stores 2126 records of 21 features extracted from cardiotocogram exams, classified by three expert obstetricians into either normal – 1655 instances (target variable 0) or suspect/pathological – 471 instances (target variable 1).

```
for each feature scheme F in [R, S, SR]:
    for each dataset D in [Diabetes, Liver, Hepatitis, Abalone, Water, Fetal]:
        for each instance i from D:
            if F == R:
                generate matrices M_LIME, M_Anchors, M_SHAP
                for each matrix Mi in [M_LIME, M_Anchors, M_SHAP]:
                    for each matrix Mj in [M_LIME, M_Anchors, M_SHAP]:
                        calculate GAME(Mi, Mj)
            if F == S or F == SR:
                generate matrices M_LIME, M_SHAP
                for each matrix Mi in [M_LIME, M_SHAP]:
                    for each matrix Mj in [M_LIME, M_SHAP]:
                        calculate GAME(Mi, Mj)

            calculate explainer confidence A
            for each explainer E:
                calculate local feature attribution weight vectors wbaraE
            calculate averaged local feature attribution weight vector wbara across all explainers E in []
        calculate averaged global feature attribution weight vector W across all instances

        split the dataset into 30% test and 70% training
        if F == R:
            for each feature overlap explanation weights method FW in [R_AVG, R_FI, R_AVG_FI, R_A, R_A_FI]:
                apply Feature weighted k-NN using FW
                calculate accuracy, F1-score, ROC-AUC score
                apply Non-weighted k-NN
                calculate accuracy, F1-score, ROC-AUC score
                compare results
        if F == S:
            for each feature overlap explanation weights method FW in [S_A, S_A_FI]:
                apply Feature weighted k-NN using FW
                calculate accuracy, F1-score, ROC-AUC score
                apply Non-weighted k-NN
                calculate accuracy, F1-score, ROC-AUC score
                compare results
        if F == SR:
            for each feature overlap explanation weights method FW in [SR_A, SR_A_FI]:
                apply Feature weighted k-NN using FW
                calculate accuracy, F1-score, ROC-AUC score
                apply Non-weighted k-NN
                calculate accuracy, F1-score, ROC-AUC score
                compare results
```

Fig. 3 Pseudocode for the algorithm that aligns the explainers and explanations and evaluates the performance of the k-NN algorithm in the two situations (aggregated feature weighted classification performance vs. non-weighted classification performance and aggregated feature weighted classification performance vs. single explanation feature weighted classification performance)

The Abalone Dataset (Nash et al. 1994) comprises 2835 recordings from male – 1528 instances (target variable 0) and female – 1307 instances (target variable 1) of abalones, described by 8 measurements of their shells (length, diameters, height, number of rings, etc.).

The Water Quality Dataset (Kadiwal 2021) contains 3276 water quality recordings – 1278 from potable sources (target variable 1) and 1998 from non-potable sources (target variable 0), characterized by 9 metrics (pH value, hardness, chloramines, sulfates, etc.).

4.3 Generating explanations

In the preprocessing step, we have identified the columns from the datasets that contained *null*, missing or NaN (not a number) values and replaced these values with the median of the valid data from the corresponding columns.

We applied the Leave-One-Out Cross-Validation procedure, that provides a robust and unbiased estimate of the model performance, by creating and evaluating the model for each instance of the dataset (Brownlee 2020). Although it is a computationally expensive method especially for complex models and large datasets, we used it because in our case, an accurate estimate of the model performance was critical for generating reliable explanations.

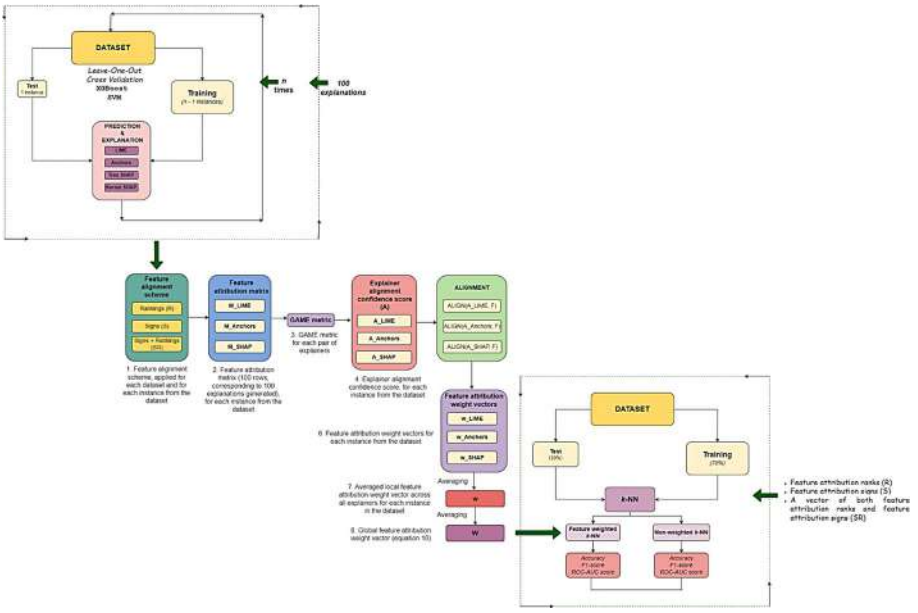

Fig. 4 Illustrative presentation of the research method in detail

Table 1 Datasets characteristics

Dataset	Number of instances	Number of target 0 instances	Number of target 1 instances	Number of features
Pima Indian Diabetes	768	500	268	8
Indian Liver Patient	583	167	416	10
Hepatitis	615	540	75	12
Fetal	2126	1655	471	21
Abalone	2835	1528	1307	8
Water Quality	3276	1998	1278	9

Each row of data from the dataset was selected as test subset and the rest of instances were assigned to the training subset. For a number of 100 times, a binary classification model was fitted on the training subset and the explanations of its predictions were saved to feature attribution vectors (Fig. 5).

For example, as the Diabetes dataset contains 768 instances, we obtained 76,800 attribution vectors for each of the 4 explainers we used (LIME, Anchors, Tree SHAP and Kernel SHAP). In the case of LIME, Tree SHAP and Kernel SHAP, for each of the 100 iterations of the algorithm that were repeated for each instance in the dataset, the feature attribution vector stored the attribution scores (the contribution of the feature to the prediction, in ascending order, starting from 1, from the most important to the least important) and sign (whether the feature has a positive or negative impact on the output – negative for driving the prediction towards the target attribute 0 and positive for routing it towards 1), for each feature (Fig. 6). The feature attribution vector looked differently for the Anchors explainers, where

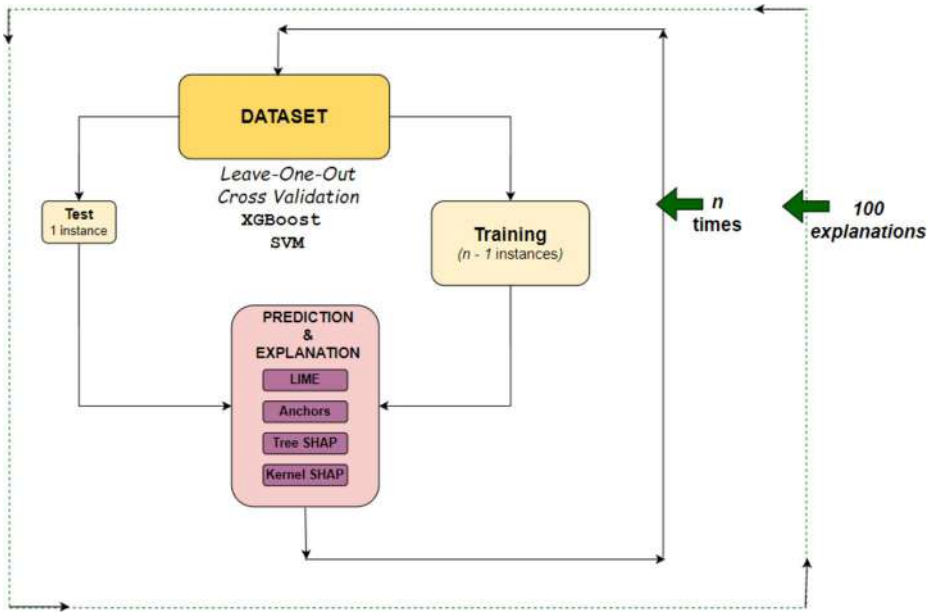

Fig. 5 Leave-one-out cross validation procedure for generating explanations

Explanation index	Feature: $feature_1$	$feature_2$	…	$feature_{m-1}$	$feature_m$
$explanation_1$	$feature\_score_{1\_1}$ $feature\_sign_{1\_1}$	$feature\_score_{1\_2}$ $feature\_sign_{1\_2}$	…	$feature\_score_{1\_m-1}$ $feature\_sign_{1\_m-1}$	$feature\_score_{1\_m}$ $feature\_sign_{1\_m}$
$explanation_2$	$feature\_score_{2\_1}$ $feature\_sign_{2\_1}$	$feature\_score_{2\_2}$ $feature\_sign_{2\_2}$	…	$feature\_score_{2\_m-1}$ $feature\_sign_{2\_m-1}$	$feature\_score_{2\_m}$ $feature\_sign_{2\_m}$
…	…	…	…	…	…
$explanation_{100}$	$feature\_score_{100\_1}$ $feature\_sign_{100\_1}$	$feature\_score_{100\_2}$ $feature\_sign_{100\_2}$	…	$feature\_score_{100\_m-1}$ $feature\_sign_{100\_m-1}$	$feature\_score_{100\_m}$ $feature\_sign_{100\_m}$

Fig. 6 Feature attribution vectors (scores and signs) for each of the 100 explanations generated for each instance from the dataset using LIME, Tree SHAP and Kernel SHAP

we considered only the attribution scores of the features mentioned in the Anchors explanation, similarly to Brughmans's method (2023). As Anchors does not compute the sign of the features, this information was absent from the Anchors feature attribution vectors.

For LIME, Anchors and Tree SHAP, we fitted an XGBoost binary classification model tuned with the $scale\_pos\_weight$ hyperparameter from the $xgboost$ Python library that

Table 2 Attribution scores for the first instance from the diabetes dataset

Dataset	Attribution scores							
	P	G	BP	ST	I	BMI	DPF	A
LIME	8	1	7	6	5	3	2	4
Tree SHAP	2	3	8	7	5	6	4	1
Kernel SHAP	5	1	7	6	3	4	8	2
Anchors		1	5		4	2	3	6

Table 3 Attribution signs for the first instance from the diabetes dataset

Dataset	Signs							
	P	G	BP	ST	I	BMI	DPF	A
LIME	negative	positive	positive	positive	positive	positive	positive	positive
Tree SHAP	negative	positive	positive	positive	positive	negative	negative	positive
Kernel SHAP	negative	negative	positive	negative	positive	negative	negative	negative

offers improved performance for binary classification models suffering from severe class imbalance. For Kernel SHAP, we chose Support Vector Machine with the radial basis kernel.

The following tables (Tables 2 and 3) show the attribution scores and signs for the first instance from the Diabetes dataset, for which all 4 algorithms predicted a positive outcome (the presence of the diabetes disease). The attribution scores highlight the contribution of the feature to the prediction, in ascending order, starting from 1, from the most important to the least important, and the sign, whether the feature has a positive or negative impact on the output – negative for driving the prediction towards the target attribute 0 and positive for routing it towards 1, for each feature:

P=Number of pregnancies.
G=Glucose level.
BP=Blood Pressure value.
ST=Skin Thickness value.
I=Insulin level.
BMI=Body Mass Index.
DPF=Diabetes Pedigree Function.
A=Age.

Anchors does not compute the sign of the features. As it can be observed from the tables, there is disagreement between explainers, especially in what concerns the attribution scores.

4.4 Aggregation by rank average

In the feature attribution aggregation strategy proposed by Pirie et al. (2023), aggregation by rank average was performed as follows: given the explanation attribution scores $S_i = [s_{ij}] \in \mathbb{R}^{n \times m}$, where n represents the number of explainers and m represents the number of features, the scores for each explainer were converted to ranks, $R_i = [r_{ij}]$, where r_{ij} signifies the attribution rank by the i -th explainer for the j -th feature.

$$r_{ij} = rank(s_{ij}) = |\{k : s_{ik} > s_{ij}\}| + 1, \; j \in [1.m] \tag{1}$$

The function $rank()$ sorts and calculates the ranks according to the order of the attribution scores. For example, if one explainer attribution scores are $s = \{2, 3, 1, 5, 4\}$, $m = 5$, considering that feature number 3 is the most important and feature number 4 is the least important, the explainer attribution ranks will be $r = \{4, 3, 5, 1, 2\}$, by assigning the highest rank to the most important feature and the lowest rank to the least important feature.

For effectively aggregating multiple explainers, feature ranks can be combined across explainers by averaging their values, resulting in an average row vector of feature weights that can be assigned to a distance-based classification algorithm such as k-NN.

$$\bar{w} = \frac{1}{n} \sum {}_{i=1}^{n} R_i \tag{2}$$

As we ran each explainer for a number of 100 times for each instance from the dataset, we further changed the notations in order to correspond to our newly adopted method.

Explanation attribution scores are given by $S_k = [s_{ij}] \in \mathbb{R}^{l \times m}$, $k \in [1.n]$, where $n = 4$ represents the number of explainers (LIME, Anchors, Tree SHAP and Kernel SHAP), m stands for the number of features of each dataset and l is the number of feature attribution vectors generated, equal to the number of instances in each dataset multiplied by 100, because each explainer was run for 100 times for each instance. For example, as the Diabetes dataset contains 768 instances, we obtained 76,800 attribution vectors ($l = 76800$).

$R_k = [r_{ij}] \in \mathbb{R}^{l \times m}$, where r_{ij} signifies the attribution rank by the i-th explanation for the j-th feature, $k \in [1.n]$, $i \in [1.l]$, $j \in [1.m]$.

$$r_{ij} = rank(s_{ij}) = |\{p : s_{ip} > s_{ij}\}| + 1, \; j \in [1.m] \tag{3}$$

The average row vector of feature ranks for each explainer k becomes:

$$\bar{w}_k = \frac{1}{l} \sum {}_{i=1}^{l} r_{ij}, \; k \in [1.n], \; j \in [1.m] \tag{4}$$

4.5 Global alignment measure (GAME)

CBR relies on the tenet that "similar problems have similar solutions". By computing the alignment between problem and solution spaces, we can quantify the degree to which the similarity hypothesis is respected in the CBR system (Raghunandan et al. 2008). As such, an efficient CBR system is characterized by a strong relationship between the knowledge containers of the problem and solution spaces. The method we used for aligning our explanations has been applied in Textual CBR (TCBR) and consists of "stacking" similar cases and features close to each other in an image derived from the case-feature matrix. Chakraborti et al. (2007) proposed a complexity measure called GAME (Global Alignment MEasure), that generates alignment scores by comparing problem and solution space clusters. The matrix of cases and features is displayed using the "case base image" metaphor, so that interesting associations and regularities are presented. Chakraborti et al. (2007) considered a set of textual cases, each case containing a set of features – definitely, in their situation, the features

were the words from the texts, and the matrix contained elements of 0 and 1 (0 if the word was not present in the text case and 1 if it was contained in the text case). The explanation matrix at this stage provides little information about the underlying patterns and complexity of the case bases. A stacking algorithm performs a twofold transformation on the case-feature matrix (an iterative process where row (case) stacking is based on case similarity and column (feature) stacking relies on feature similarity), so that similar cases and features are grouped together. Weighted similarity makes sure that more recently stacked rows/columns have a higher contribution to the decision of which next row/column will be stacked.

***Stack_Rows*(*M*)**

Input: Case-Feature matrix M

Output: Case-Feature matrix M_R (M stacked by rows)

Method:

Instantiate first row of M_R to first row of M
for $i = 2$ to $noOfRows$
for $j = i$ to $noOfRows$
$wsim_j = 0$
for $k = 1$ to $i - 1$
$wsim_j = wsim_j + (1 / (i - k)) * sim(c_k, c_j)$
end
end
choose j that maximizes $wsim_j$ and swap rows i and j
end
Return M_R

***Stack_Columns*(M_R)**

Input: Case-Feature matrix M_R generated by *Stack_Rows*(M)

Output: Case-Feature matrix M_C (M_R stacked by columns)

Method:

Instantiate first row of M_C to first row of M_R
for $i = 2$ to $noOfCols$
for $j = i$ to $noOfCols$
$wsim_j = 0$
for $k = 1$ to $i - 1$
$wsim_j = wsim_j + (1 / (i - k)) * sim(f_k, f_j)$
end
end
choose j that maximizes $wsim_j$ and swap columns i and j
end
Return M_C

$sim(c_k, c_j)$ represents the cosine similarity function between cases c_k and c_j and $sim(f_k, f_j)$ stands for the cosine similarity function between features f_k and f_j.

The result of this algorithm is a clustered image where feature associations and case similarity patterns are exposed.

For globally measuring the alignment between problem and solution spaces, two case-feature matrices need to be stacked using the twofold transformation presented above (one from the problem side and one from the solution side). The matrix M_p obtained by stacking the problem side contains the best ordering of the cases and features from the problem space and the matrix M_s resulted by stacking the solution side, displays the most effective arrangement in terms of underlying similarity structure from the solution side. For measuring alignment, we need to compare the ordering of cases in M_p and M_s. A third matrix, M_{sp}, is generated by stacking the M_s matrix using the case ordering from the problem space, as retrieved from M_p. The degree of similarity between the case ordering from M_{sp} and the case ordering from M_s is an indicator of alignment.

The average similarity of a case to its neighbors (calculated as a weighted sum of similarity values to the k previously designated cases with exponentially decaying weight) in the clustered matrix M_C has the following formula:

$$Sim\,(c_i) = \sum_{j=1}^{k} sim\,(c_i,\ c_{i-j}) * \frac{1}{j} \tag{5}$$

The average similarity of the clustered matrix M_C is:

$$Sim\,(M_C) = \frac{\sum_{i=2}^{N} Sim\,(c_i)}{N-1} \tag{6}$$

N denotes the number of cases in M_C

This formula is applied to matrices M_s and M_{sp} to obtain $Sim\,(M_s)$ and $Sim\,(M_{sp})$. Global Alignment Measure is therefore calculated as:

$$GAME = \frac{Sim\,(M_{sp})}{Sim\,(M_s)} \tag{7}$$

In the CBR systems where there is a strong alignment between problem and solution spaces, the best problem side ordering should be similar to the solution side ordering and the GAME metric should have a value very close to 1. Contrastingly, for the datasets that are weakly aligned, GAME quantifies a value far from 1 (Raghunandan et al. 2008).

4.6 GAME in our aggregation strategy

In our approach, the cases from the matrix were represented by explanations and the features were represented by either: **1**. feature attribution ranks (**R**) obtained from our explanations using the method described in subchapter 4.4. **2**. feature attribution signs (**S**) and **3**. a vector of both feature attribution ranks and feature attribution signs (**SR**).

For the **R** feature scheme (feature attribution ranks), for each dataset and for each instance from the dataset, we generated the matrices **M_Anchors**, **M_LIME**, and **M_SHAP**. These matrices had a number of 100 rows, corresponding to the number of explanations produced by each explainer for every instance from the dataset in the iterative explanation generation

process previously described in subchapter 4.3. Tree SHAP was stable and consistent across the 100 explanations generated for each instance from the datasets, meaning that all 100 explanations were identical. The same situation happened with Kernel SHAP. Kundu and Hoque (2023) made a similar observation - two identical instances always have the same explanation generated by SHAP, but this is not the case for either LIME or Anchors, as these methods generate potentially unstable explanations. Given this fact, when formulating the **M_SHAP** matrix, we populated it with 50 explanations from Tree SHAP and 50 explanations from Kernel SHAP. For each instance from the dataset, we aligned the matrices **M_Anchors**, **M_LIME** and **M_SHAP** in a pairwise fashion using the cluster-based Global Alignment approach and then we calculated the GAME metric for each pair (Anchors-Anchors, Anchors-LIME, Anchors-SHAP, LIME-Anchors, LIME-LIME, LIME-SHAP, SHAP-Anchors, SHAP-LIME, SHAP-SHAP). Obviously, the value of the GAME metric when the explainers were identical (Anchors-Anchors, LIME-LIME, SHAP-SHAP) was 1.

For each dataset, the GAME scores have been stored in a .xlsx file having 9 columns (for each GAME value corresponding to the pairwise combinations of explainers - Anchors-Anchors, Anchors-LIME, Anchors-SHAP, LIME-Anchors, LIME-LIME, LIME-SHAP, SHAP-Anchors, SHAP-LIME, SHAP-SHAP) and a number of rows equal to the number of instances from the dataset.

As Anchors does not generate the sign of the feature contribution in the explanation, it has been excluded from the schemes **S** and **SR**. In this situation, we aligned only the matrices **M_LIME** and **M_SHAP** and we calculated the GAME metric for the pairs LIME-LIME, LIME-SHAP, SHAP-LIME and SHAP-SHAP. For each dataset, the GAME scores have been stored in a .xlsx file with 4 columns. In our alignment strategy, the GAME measurement employed the cosine similarity function, as did in the research of Chakraborti et al. (2007) on textual case bases. Cosine similarity presents several advantages, such as being scale-invariant and robust in capturing pattern similarities.

Let $M = [m_{ij}] \in \mathbb{R}^{n \times n}$, a matrix representing the pairwise alignment relationships (GAME metrics) between the explainers. m_{ij} is the GAME score between explainer i and explainer j.

$n = 3$ for the scheme **R** and $n = 2$ for the schemes **S** and **SR**. Pirie et al. (Pirie et al. 2023) proposed the explainer confidence vector A = $[A_i] \in \mathbb{R}^n$, where:

$$A_i = \begin{cases} \frac{1}{n}\sum_{j=1}^{n} m_{ij}, \; if \; M \; is \; symmetric \\ \frac{1}{2n}\sum_{j=1}^{n} (m_{ij} + m_{ji}), \; \text{otherwise} \end{cases} \tag{8}$$

A_i is the confidence score for each explainer i. It is used to influence the degree of importance given to each feature attribution vector that contains: ranks (for alignment scheme **R**), signs (for alignment scheme **S**) and both signs and ranks (for alignment scheme **SR**).

For each alignment scheme **R**, **S** and **SR**, for each dataset, for each instance from the dataset, we calculated the explainer confidence vector **A** and then, for each of the 100 explanations, we computed the consensus feature attribution weight vector for every explainer i.

$$\bar{w}_i^c = \frac{\sum_{k=1}^{n} (A_k * F_i^c)}{\sum_{k=1}^{n} A_k} \tag{9}$$

c represents the index of the current explanation from the 100 explanations generated for each dataset $c \in [1;100]$

i represents the index of the current explainer, $i \in [1;n]$, $n = 3$ for the scheme **R** and $n = 2$ for the schemes **S** and **SR**.

A_k is the confidence score for the k-th explainer

$\boldsymbol{F_i^c}$ is a row vector containing:

- For scheme **R**: feature attribution ranks obtained from our explanations using the method described in subchapter 4.4.
- For scheme **S**: feature attribution signs.
- For scheme **SR**: both feature attribution ranks and feature attribution signs.

For each instance, the feature attribution weight vectors $\bar{\boldsymbol{w}}_{\boldsymbol{i}}^{\boldsymbol{c}}$ have been averaged, resulting in local feature attribution weight vectors $\bar{\boldsymbol{w}}_{\boldsymbol{i}}$ for each explainer i.

By averaging $\bar{\boldsymbol{w}}_i$, we obtained $\overline{\boldsymbol{w}}$, which is the averaged local feature attribution weight vector across all explainers for each instance in the dataset.

If N is the number of instances from the dataset, the global feature attribution weight vector $\boldsymbol{W}$ that averages the local weight vectors across all instances is calculated as follows:

$$W = \frac{1}{N}\sum_{i=1}^{N}\overline{w} \tag{10}$$

The aggregation strategy workflow is described in Fig. 7.

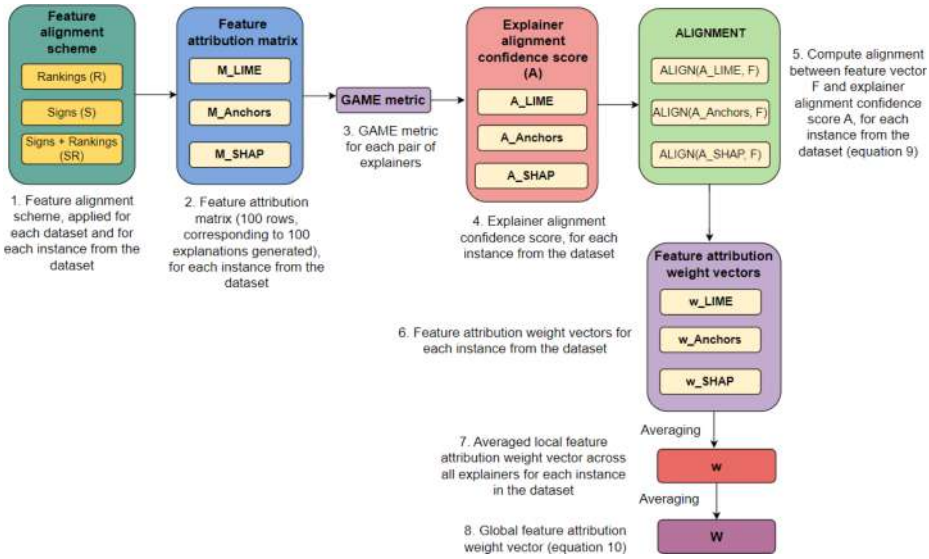

Fig. 7 Aggregation strategy workflow

5 Method evaluation

5.1 Aggregated feature weighted classification performance vs. non-weighted classification performance

As Pirie et al. (2023) assumed, explanation aggregation captures feature significance by arriving at a consensus between explainers, which in consequence can improve classification performance. We weighted the feature space of a k-NN classifier with the aligned importance of each feature (for each alignment scheme **R**, **S**, **SR**) and compared the prediction accuracy against a non-weighted k-NN classifier (Fig. 8). Feature weighted k-NN (FWk-NN) gives each feature a different weight, so that the most important ones will have a higher contribution to the prediction (Chen and Hao 2017). This stems from the fact that k-NN is inherently sensitive to irrelevant features.

We tested the k-NN $(k = 5)$ algorithm using the following feature overlap explanation weights:

- R_AVG – the mean of rankings. When computing R_AVG, in Eq. (9), the explainer confidence vector **A** contained only values of 1.
- R_FI – the mean of feature importances as obtained from the XGBoost classifier using the $feature\_importances\_$ function from the $xgboost$ Python library.
- R_AVG_FI – the mean of rankings multiplied by the mean of feature importances.
- R_A – the mean of rank alignments obtained from the Global Alignment Measurement described in subchapter 4.6.
- R_A_FI – the mean of rank alignments obtained from the Global Alignment Measurement described in subchapter 4.6., multiplied by the mean of feature importances as obtained from the XGBoost classifier using the $feature\_importances\_$ function from

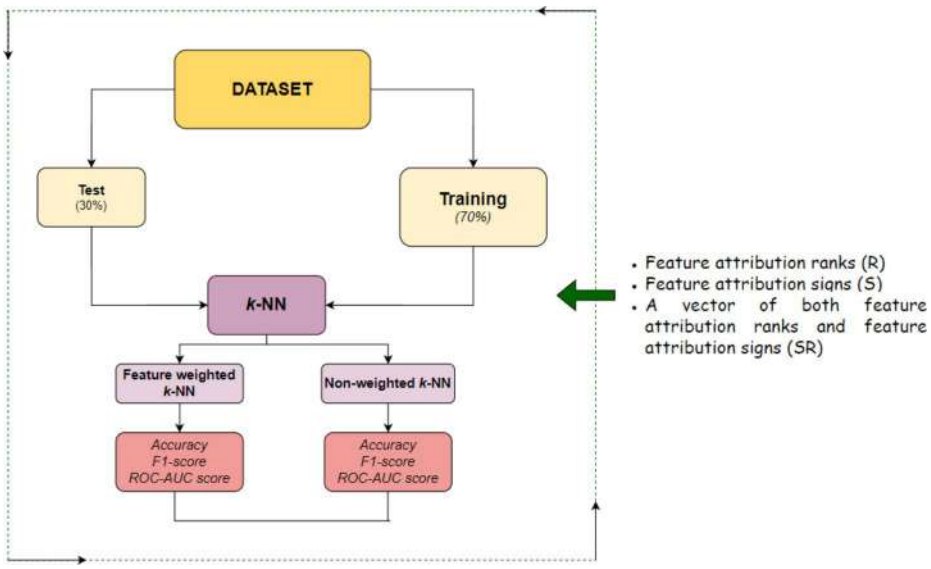

Fig. 8 Evaluation of the feature weighted and non-weighted k-NN algorithms

the $xgboost$ Python library
- S_A – the mean of sign alignments obtained from the Global Alignment Measurement described in subchapter 4.6.
- S_A_FI - the mean of sign alignments obtained from the Global Alignment Measurement described in subchapter 4.6., multiplied by the mean of feature importances as obtained from the XGBoost classifier using the $feature\_importances\_$ function from the $xgboost$ Python library
- SR_A – the mean of the vector containing rank and sign alignments obtained from the Global Alignment Measurement described in subchapter 4.6.
- SR_A_FI - the mean of the vector containing rank and sign alignments obtained from the Global Alignment Measurement described in, multiplied by the mean of feature importances as obtained from the XGBoost classifier using the $feature\_importances\_$ function from the $xgboost$ Python library.

Given the explanation weights $\overline{\boldsymbol{w}}$ from any of these feature overlap methods, we calculated the weighted Euclidean distance between two instances, x and y, as follows:

$$dist\,(x,\ y) = \sqrt{\sum_{i=1}^{m} \overline{w}(x_i - y_i)^2} \tag{11}$$

where m represents the number of features of each instance.

For the non-weighted k-NN, the $\overline{\boldsymbol{w}}$ vector was filled with elements of 1.

We ran the weighted and the non-weighted k-NN classifiers for a number of 50 times and then we averaged the results. Performance was quantified using accuracy and two additional metrics suitable for imbalanced datasets, namely F1-score and ROC-AUC score. The Global Alignment framework is presented in Fig. 9.

5.1.1 Comparison between the averaged metrics of the weighted k-NN predictions using each of the 9 feature overlap explanation weights to the predictions of the non-weighted k-NN classifier

We applied the Wilcoxon signed-rank test, with a significance level of 5% to verify if the averaged accuracies, F1-scores and ROC-AUC scores of the weighted k-NN predictions using each of the 9 feature overlap explanation weights were different from the predictions of the non-weighted k-NN classifier. To counteract the problem of multiple comparisons and the occurrence of Type 1 errors, we employed the Holm-Bonferroni method that adjusts the rejection criterion for each of the individual hypotheses. The results are presented in Tables 4, 5, 6 in the Appendix. The null hypothesis states that there is no difference (in terms of central tendency) between the two groups in the population. If the *p*-value is less than or equal to the significance level, the decision is to reject the null hypothesis, concluding that there is a difference between the groups. The cells not containing percentages signify that the predictions of the weighted k-NN algorithm using the respective feature overlap explanation weights were not significantly different from the predictions of the non-weighted k-NN classifier. The values in bold represent the maximum performance metric for each of the datasets.

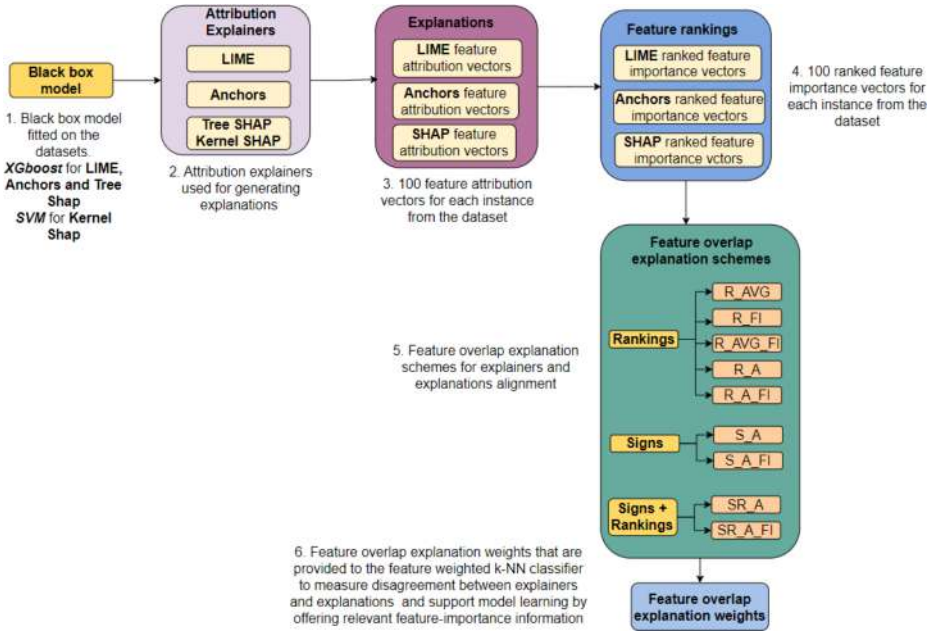

Fig. 9 Global alignment framework

The averaged accuracies, F1-scores and ROC-AUC scores of the weighted k-NN algorithm were higher than those of the non-weighted k-NN classifier for 5 out of 6 datasets. The only exception was the Abalone dataset, which, in fact, does not provide satisfactory classification results at all. Accuracies of around 50% and F1-scores of around 40% are similar to a random guess for a balanced dataset such as this one. Table 7 in the Appendix and Fig. 10 present the increase (in %) from the baseline non-weighted k-NN performance metric scores (accuracy, F1-score, ROC-AUC score) to the averaged (averaged across feature overlap methods) weighted k-NN performance metric scores, for each dataset.

The highest increase from the baseline non-weighted k-NN as a result of applying feature overlap methods (rankings averaging, feature importance averaging, the explainer aggregation strategy across multiple explainers and explanations using the Global Alignment Measurement algorithm for rankings and signs) has been achieved for the Hepatitis dataset (a notable improvement of 16.49% for the F1-score, from 55.01 to 74.7% and of 10.19% for the ROC-AUC score, from 69.59 to 79.78%), for the Diabetes dataset (improvement of 4.93% for the F1-score, from 55.74 to 60.67% and of 3.08% for the ROC-AUC score, from 67.34 to 70.42%) and for the Fetal dataset (improvement of 3.09% for the F1-score, from 79.95 to 83.04% and of 2.24% for the ROC-AUC score, from 85.88 to 88.13%).

The feature overlap explanation methods that contributed to the highest accuracies, F1-scores and ROC-AUC scores compared to the non-weighted k-NN for all datasets were: **R_A_FI** for the Diabetes, Water and Fetal datasets, **R_AVG** for Liver and Hepatitis datasets and **R_FI** for the Abalone dataset. The difference in prediction performance between the most efficient weighted k-NN model and the baseline non-weighted k-NN classifier is presented in Table 8 in the Appendix and Fig. 11.

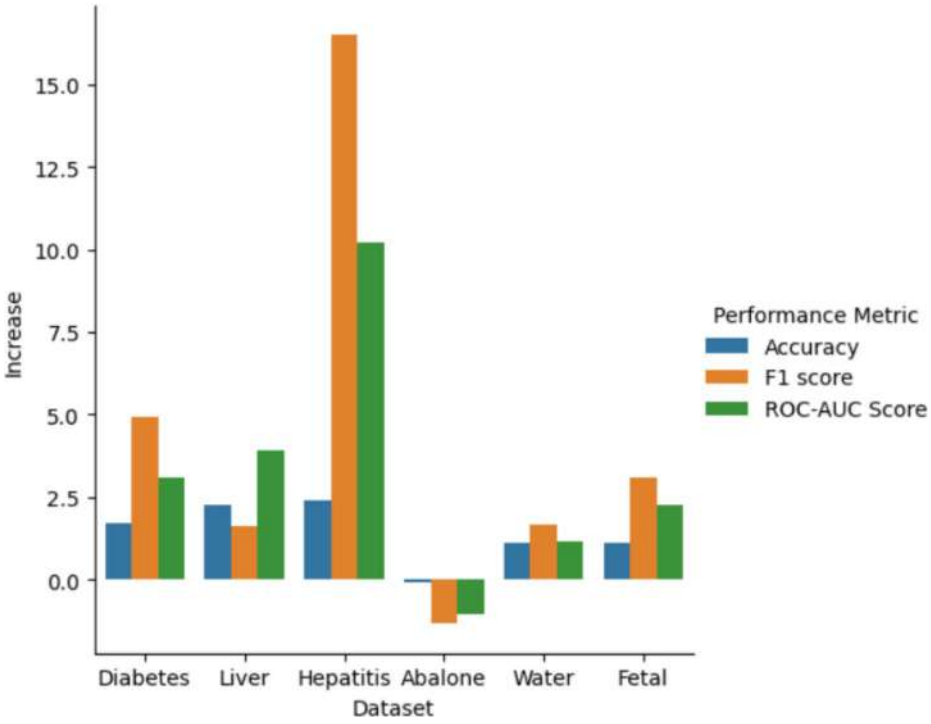

Fig. 10 Plot of the increase (in %) from non-weighted k -NN compared to averaged weighted k -NN performance metrics for each dataset

The highest increase from the baseline non-weighted k -NN predictions has been obtained by applying the R_AVG_FI feature overlap explanation weight vector (the mean of rankings multiplied by the mean of feature importances) to the weighted k -NN binary classifier in the case of the Hepatitis dataset (an increase of 24.05% for the F1-score, from 55.01 to 79.07% and of 15.43% for the ROC-AUC score, from 69.59 to 85.02%). For the Diabetes dataset, the highest improvement has been achieved by applying the R_A_FI feature overlap explanation weight (the mean of rank alignments obtained using the Global Alignment Measurement approach, multiplied by the mean of feature importances) – an increase of 7.05% for the F1-score, from 55.74 to 62.79% and an increase of 4.59% for the ROC-AUC score, from 67.34 to 71.93%. The R_A_FI feature overlap explanation weight contributed to the highest difference from the baseline non-weighted *k*-NN for the Fetal dataset - an increase of 4.86% for the F1-score, from 79.95 to 84.82% and an increase of 3.46% for the ROC-AUC score, from 85.88 to 89.35%. As Pirie et al. (2023) suggested, a stable or increased performance score of the feature-agreed weighed k -NN algorithm, compared to the non-weighted k -NN version, is an indicator of good agreement between explainers. In our approach, by applying the Global Alignment Measurement, we demonstrated that **R_AVG_FI** and **R_A_FI** are the feature overlap explanation weight methods with the highest impact on explainers and explanations consensus. They increase classification performance (especially F1-score and ROC-AUC score) and support model learning by offering relevant feature-importance information.

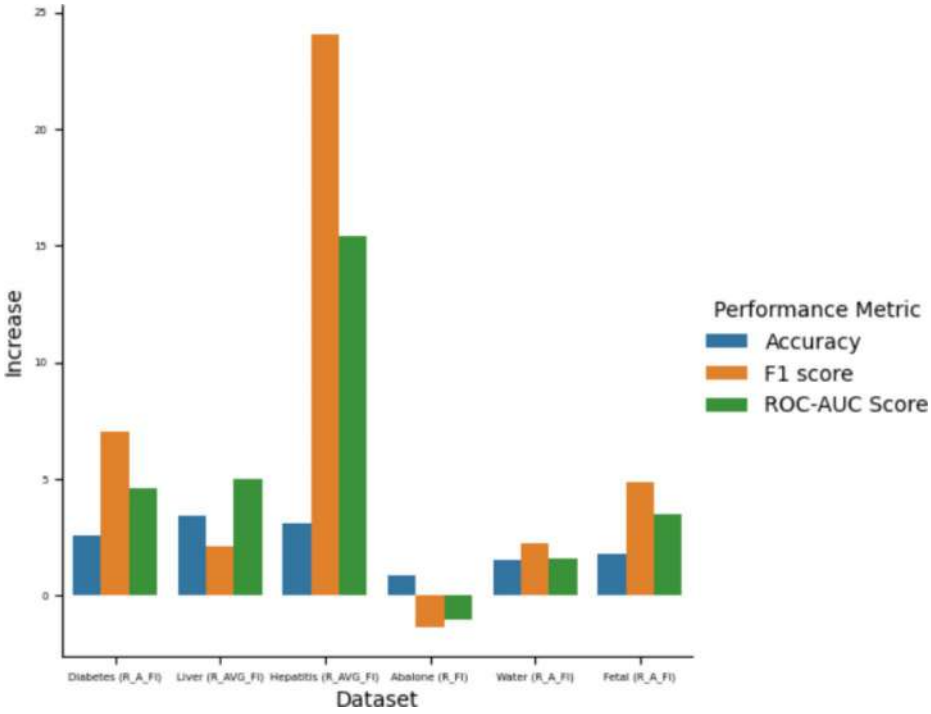

Fig. 11 Plot of increase (in %) from the best performing feature overlap weighted k -NN model compared to the non-weighted k -NN for each dataset

5.1.2 Comparison between the average increase (in %) of a feature overlap explanation weight method performance to the rest of the feature overlap explanation weight methods

We repeatedly computed the Wilcoxon signed-rank test to compare the accuracies, F1-scores and ROC-AUC scores of the weighted k -NN classifier using each of the 9 feature overlap explanation weights. As we performed multiple tests, we applied the Holm-Bonferroni correction method, that is fairly simple to implement and more powerful than the single-step Bonferroni. For each performance metric (accuracy, F1-score and ROC-AUC score), for each of the 9 feature overlap explanation weight methods, we obtained the feature overlap explanation weight methods that were significantly different from the one tested.

The feature overlap explanation weight methods that overcome the others, in terms of prediction performance for all three metrics, are **R_AVG_FI** and **R_A_FI** (all the cells from their corresponding rows in Tables 9, 10, 11 in the Appendix contain dataset names that are colored in green). We also notice that their classification results are similar – R_AVG_FI and R_A_FI are not statistically different for any of the 6 datasets. As a result, we can certainly infer that R_AVG_FI (the mean of rankings multiplied by the mean of feature importances) and R_A_FI (the mean of rank alignments obtained from the Global Alignment Measurement described in subchapter 4.6, multiplied by the mean of feature importances as obtained from the XGBoost classifier using the $feature\_importances\_$ function from the xgboost

Python library) are the feature overlap explanation weight approaches that contribute the most to: reaching an agreement between explainers and their explanations, offering support in model learning by bringing in reliable feature-importance knowledge and enhancing the performance of the weighted k -NN classifier.

Table 12 in the Appendix and Fig. 12 present the average increase (in %) of a feature overlap explanation weight method performance (accuracy, F1-score and ROC-AUC score), compared to the rest of the feature overlap explanation weight methods, for all datasets.

R_AVG_FI and R_A_FI have a similar increase in classification performance (R_AVG_FI – 1.57% and R_A_FI – 1.63% for accuracy, R_AVG_FI – 4.03% and R_A_FI – 4.34% for F1-score, R_AVG_FI – 3.56% and R_A_FI – 3.2% for ROC-AUC score), compared to the rest of feature overlap explanation weight methods. They are followed by R_FI (0.78%), S_A_FI (0.74%), SR_A_FI (0.46%), R_A (-1.06%), R_AVG (-1.15%), S_A (-2.5%) and SR_A (-3.11%) – these values are the averages of the 3 performance metrics. As observed, the feature overlap explanation weight methods that incorporate a contribution of feature importances perform better compared to the others (R_AVG_FI and R_A_FI).

Table 13 in the Appendix presents the accuracy of the *k*-NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the accuracy of *k*-NN when employing the other 8 feature overlap weights, for each dataset. Table 14 in the Appendix presents the accuracy of the *k*-NN classifier when using the R_A_FI feature overlap weights (in

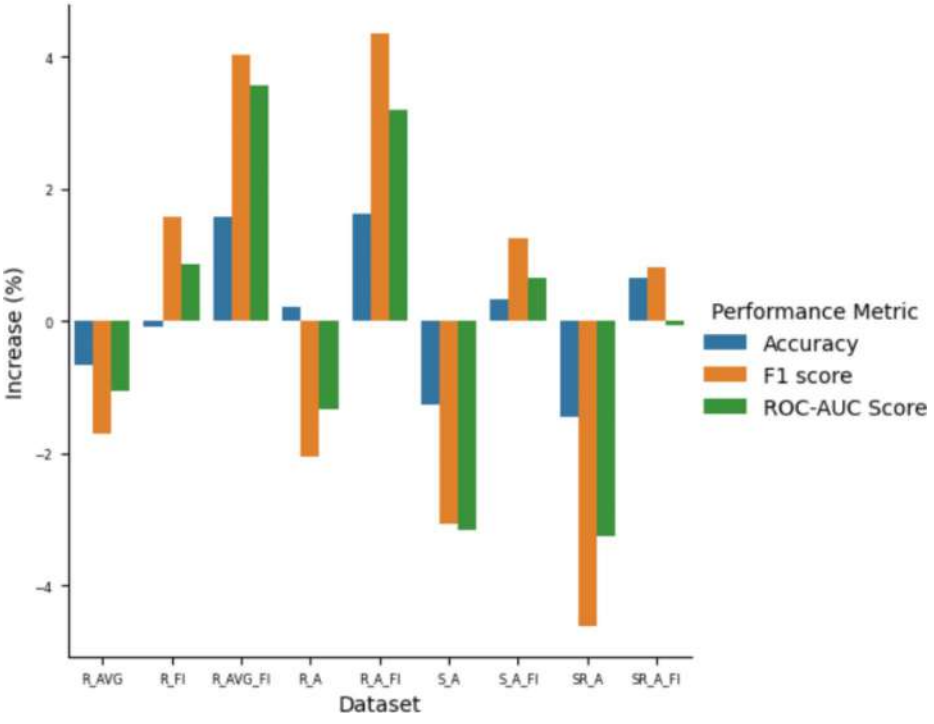

Fig. 12 Plot of the average increase (in %) of a feature overlap explanation weight method performance (accuracy, F1-score and ROC-AUC score), compared to the rest of the feature overlap explanation weight methods, for all datasets

bold) and also the accuracy of *k*-NN when employing the other 8 feature overlap weights, for each dataset.

Figure 13 presents the accuracy of the k -NN classifier when using the R_A_FI feature overlap weights (in bold) and also the accuracy of k -NN when employing the other 8 feature overlap weights, for each dataset.

Table 15 in the Appendix presents the F1-score of the *k*-NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the F1-score of *k*-NN when employing the other 8 feature overlap weights, for each dataset. Table 16 in the Appendix presents the F1-score of the *k*-NN classifier when using the R_A_FI feature overlap weights (in bold) and also the F1-score of *k*-NN when employing the other 8 feature overlap weights, for each dataset.

Figure 14 presents the F1-score of the k -NN classifier when using the R_AVG_FI and R_A_FI feature overlap weights (in bold) and also the F1-score of k -NN when employing the other 8 feature overlap weights, for each dataset.

Table 17 in the Appendix presents the ROC-AUC score of the *k*-NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the ROC-AUC score of *k*-NN when employing the other 8 feature overlap weights, for each dataset. Table 18 in the Appendix presents the ROC-AUC score of the *k*-NN classifier when using the R_A_FI feature overlap weights (in bold) and also the ROC-AUC score of *k*-NN when employing the other 8 feature overlap weights, for each dataset.

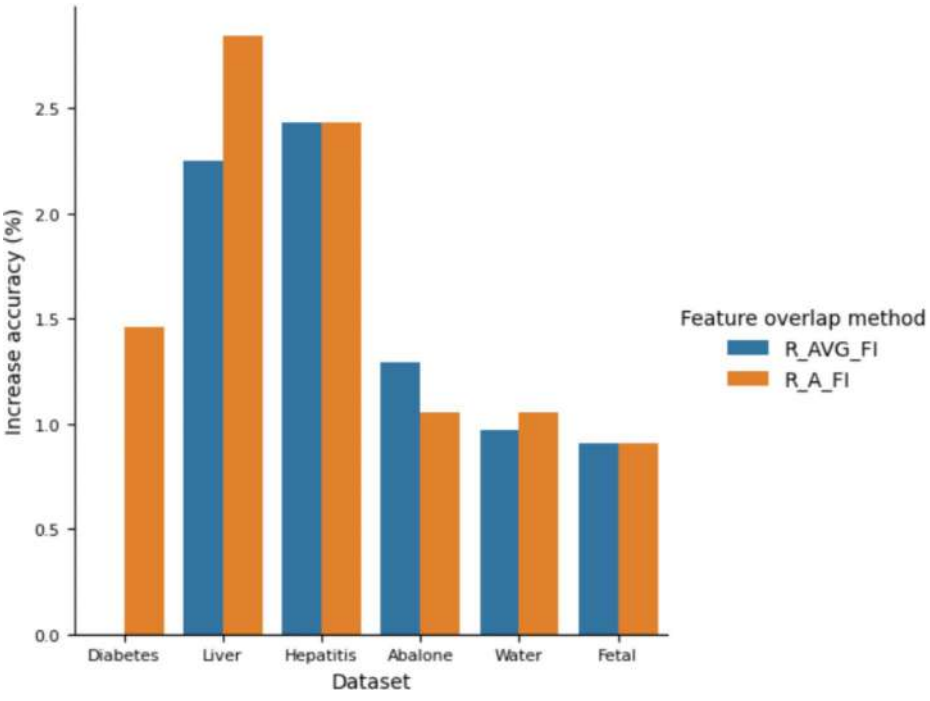

Fig. 13 Increase (in %) for the accuracy of the k -NN classifier when comparing R_AVG_FI and R_A_FI with the average accuracy of the other feature overlap methods

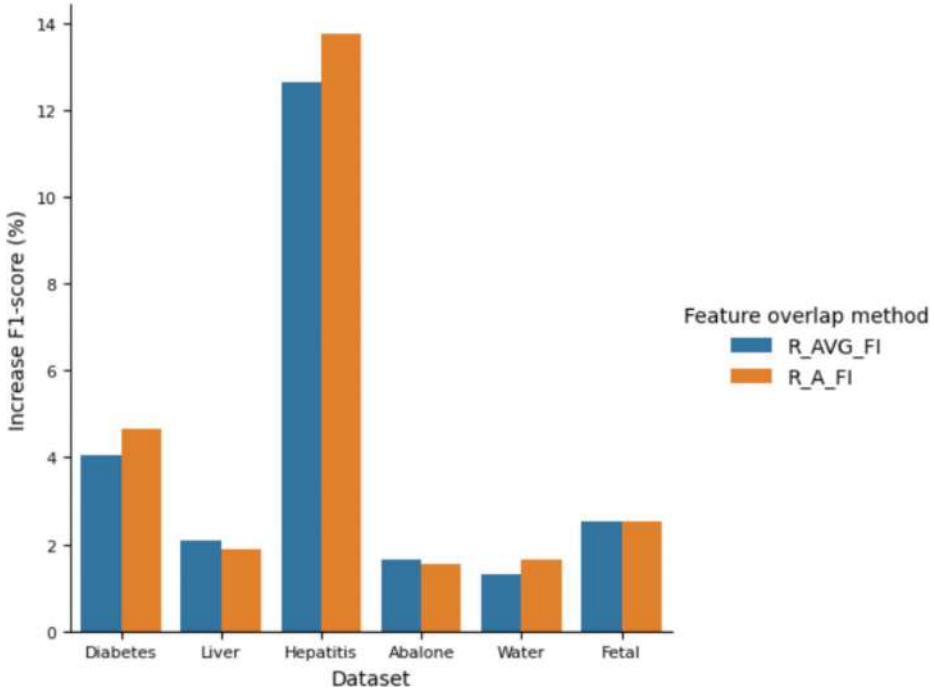

Fig. 14 Increase (in %) for the F1-score of the k -NN classifier when comparing R_AVG_FI and R_A_FI with the average F1-score of the other feature overlap methods

Figure 15 presents the ROC-AUC score of the k -NN classifier when using the R_AVG_FI and R_A_FI feature overlap weights (in bold) and also the ROC-AUC score of k -NN when employing the other 8 feature overlap weights, for each dataset.

5.2 Aggregated feature weighted classification performance vs. single explanation feature weighted classification performance

We compared the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), for each feature alignment scheme R, S and SR. Actually, for the Anchors explanation method we had only the R feature alignment scheme, as Anchors does not compute signs in its explanations.

Tables 19, 20, 21, 22, 23, 24, 25 in the Appendix present the increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by a single explanation method to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset.

For the Diabetes dataset, the highest averaged increase has been obtained from the LIME – R weights (accuracy – 2.84%, F1-score – 7.67%, ROC-AUC score – 4.85%) and the lowest averaged increase has been obtained from the Anchors – R weights for accuracy (0.33%) and from the SHAP – SR weights for F1-score – 0.31% and ROC-AUC score 0.36%.

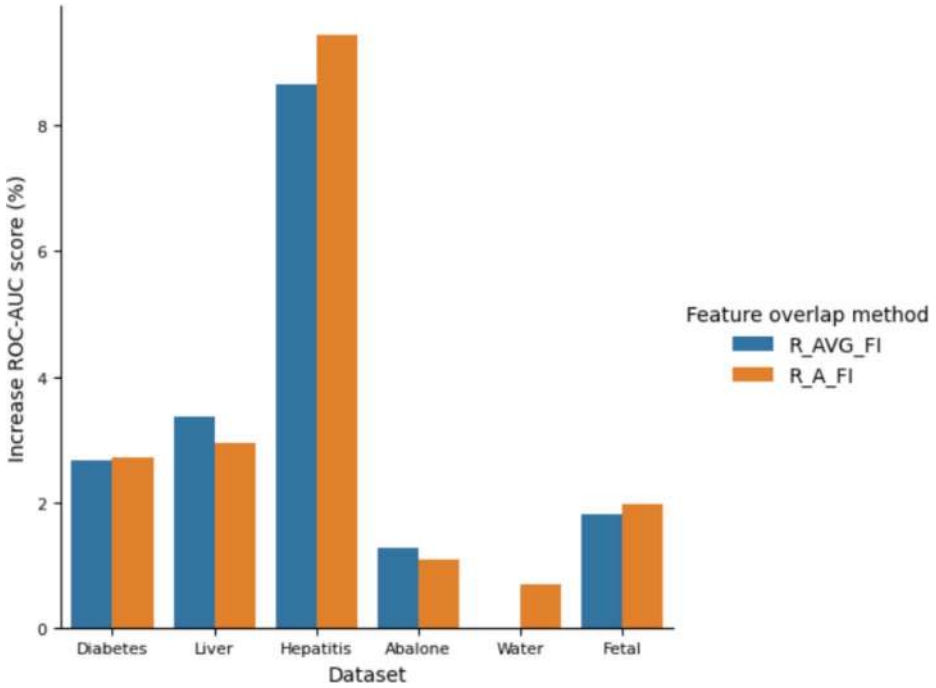

Fig. 15 Increase (in %) for the ROC-AUC score of the k -NN classifier when comparing R_AVG_FI and R_A_FI with the average ROC-AUC score of the other feature overlap methods

For the Liver dataset, the highest averaged increase has been obtained from the SHAP – R weights (accuracy – 1.52%, F1-score – 0.82%, ROC-AUC score – 2.63%).

For the Hepatitis dataset, the highest averaged increase has been obtained from the SHAP – R weights (accuracy – 2.1%, F1-score – 16.54%, ROC-AUC score – 10.19%).

For the Abalone dataset, the highest averaged increase has been obtained from the SHAP – R weights (accuracy – 1.2%, F1-score – 1.48%, ROC-AUC score – 1.23%).

For the Water dataset, the highest averaged increase has been obtained as follows: from the SHAP – S weights for the accuracy metric (2.23%), from the LIME – R weights for the F1-score metric (4.54%) and from the LIME – SR weights for the ROC-AUC score (2.97%).

For the Fetal dataset, the highest averaged increase has been obtained as follows: from the SHAP – R weights for the accuracy metric (1.68%), from the LIME – SR weights for the F1-score metric (4.24%) and from the LIME – R weights for the ROC-AUC score (2.67%).

The highest averaged increase is therefore obtained for the R feature alignment scheme, which uses only feature attribution ranks, for either the LIME and SHAP explainers.

Table 26 in the Appendix and Fig. 16 present the increase (in %) from the averaged (averaged across explanation methods) weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by a single explanation method to the averaged (averaged across feature overlap methods) weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset.

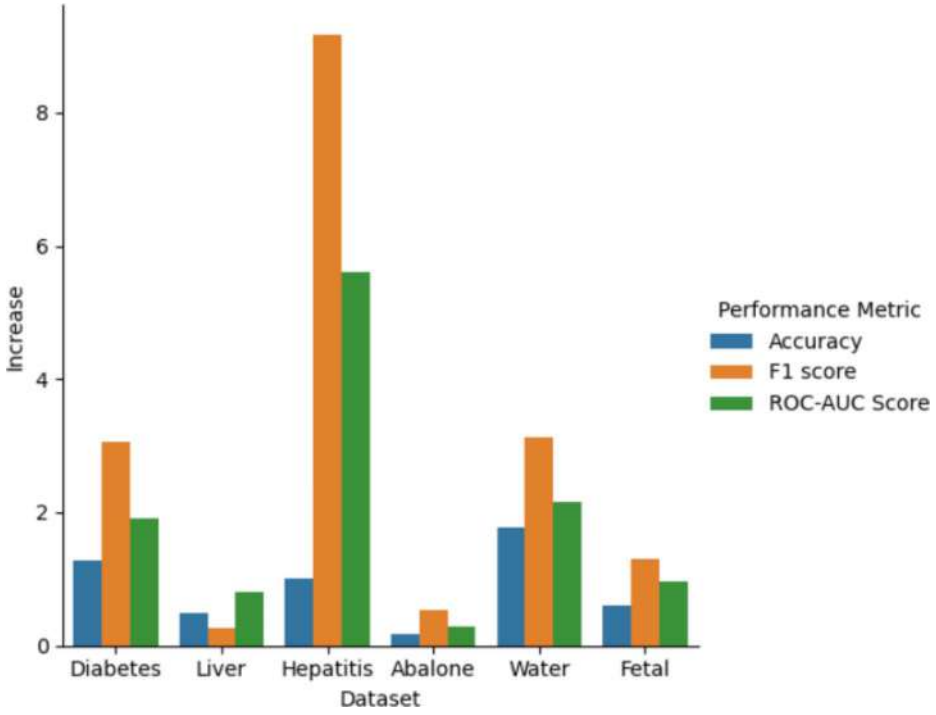

Fig. 16 Increase (in %) (averaged single explanation method weighted k -NN compared to averaged feature aggregated weighted k -NN)

The results show a positive increase from the averaged (averaged across explanation methods) weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by a single explanation method to the averaged (averaged across feature overlap methods) weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset, demonstrating that by weighting the feature space of the k -NN classifier with the agreed feature overlap explanation weights we can obtain better results than by using only single explanations weights.

The highest averaged increase in accuracy has been obtained for the Water dataset (1.78%). The highest averaged increase for F1-score has been obtained for the Hepatitis dataset (9.16%) and for ROC-AUC score, for the Hepatitis dataset as well (5.61%).

6 Discussion

After weighting the latent space of a k -NN classifier using the weights generated by the nine feature overlap explanation weight methods, we observed an increase in performance, compared to the baseline non-weighted k -NN, for 5 out of the 6 datasets. The only dataset that had a decrease in performance was Abalone, a balanced dataset – balance ratio of 1:1.16 (a number of 1307 target variables of 1 and 1528 target variables of 0) with 8 features. The maximum accuracy for this dataset was 54.47%, F1-score of 46.44% and ROC-AUC score

of 52.07% when using the R_FI feature overlap method. The non-weighted k-NN algorithm evaluated accuracy to 53.62%, F1-score to 47.78% and ROC-AUC score to 53.13%. These values of performance are very weak, signaling that the Abalone dataset has a sub-optimal classification characteristic that cannot be improved by any of the proposed feature aggregation strategies.

The Hepatitis dataset, which is severely imbalanced (a balance ratio of 1:7.2), reached an accuracy of 95.54%, F1-score of 79.07% and ROC-AUC score of 85.02% when using the R_AVG_FI feature explanation overlap method, an increase of 3.1% for accuracy, 24.05% for F1-score and 15.43% for ROC-AUC score compared to the baseline non-weighted k-NN classifier. On the second and on the third places were the Diabetes and Fetal datasets, which are also imbalanced. The Diabetes dataset has a balance ratio of 1:1.86 and the Fetal dataset, of 1:3.51. For them, the best classification performance has been obtained when using the R_A_FI feature overlap method. They were followed by the Liver dataset (balance ratio of 1:2.5) and the Water dataset, that is quite balanced (balance ratio of 1:1.56). For the Liver dataset, the best performing feature overlap method was R_AVG_FI, while for the Water dataset, R_A_FI. The performance metrics of the majority of the 9 feature overlap explanation weight methods were statistically different from the performance of the baseline non-weighted k-NN classifier, these values were similar to each other, but the feature overlap explanation weight strategies that outperformed the rest were definitely R_AVG_FI and R_A_FI.

R_AVG_FI and R_A_FI performed better than the non-weighted k-NN for 5 out of the 6 datasets (except for Abalone) - accuracy (on average for R_AVG_FI with 2.32% and for R_A_FI with 2.38%), F1-score (on average for R_AVG_FI with 7.81% and for R_A_FI with 7.94%) and for ROC-AUC score (on average for R_AVG_FI with 5.76% and for R_A_FI with 5.71%).

S_A_FI and SR_A_FI performed better than the non-weighted k-NN for 5 out of the 6 datasets (except for Abalone) for accuracy (on average for S_A_FI with 1.42% and for SR_A_FI with 1.61%) and F1-score (on average for S_A_FI with 5.65% and for SR_A_FI with 5.85%) and for 4 out of the 6 datasets (except for Abalone and Liver) for ROC-AUC score (on average for S_A_FI with 4.55% and for SR_A_FI with 4.38%).

R_AVG_FI had an increase of 1.57% for accuracy, 4.03% for F1-score, 3.56% for ROC-AUC score compared to the average of the other feature overlap methods. This increase is statistically similar in terms of all classification metrics for R_A_FI: an increase of 1.63% for accuracy, 4.34% for F1-score, 3.2% for ROC-AUC score, compared to the average of the other feature overlap methods. The increase in performance when using the R_AVG_FI and the R_A_FI feature overlap explanation weight strategies compared to the rest of the feature overlap explanation weight methods tested was achieved for all datasets, for all 3 evaluation metrics (accuracy, F1-score and ROC_AUC score). They were followed by S_A_FI (an increase of 0.33% for accuracy, 1.24% for F1-score, 0.66% for ROC-AUC score compared to the average of the other feature overlap methods) and SR_A_FI (an increase of 0.65% for accuracy, 0.82% for F1-score, -0.07% for ROC-AUC score compared to the average of the other feature overlap methods).

From the results presented, we highlight the significance of feature importances when hybridizing the feature overlap explanation weight methods. Although previous research

assumed that each feature contributed equally to classification, Chen and Hao (2017) motivated that "some features are closely relevant to the classification, some are trivial relevant, and others are irrelevant". Feature importances, when applied properly into the classifier, without being dominated by trivial relevant or irrelevant features, can improve its robustness and quality performance. Feature weighted k -NN presented significant improvement and good performance for Chinese stock market indices prediction compared to other models in the research of Chen and Hao (2017). The criterion of giving each feature a weight value corresponding to its information gain proved to be valid in our aggregation strategy across explainers and explanations as well. As detailed above, the feature overlap explanation methods that performed better than the rest and also achieved a higher classification performance when applied to the weighted k -NN algorithm, compared to the non-weighted k -NN, were the ones that incorporated a contribution of feature importances: R_AVG_FI (the mean of rankings multiplied by the mean of feature importances), R_A_FI (the mean of rank alignments obtained from the Global Alignment Measurement described in subchapter 4.6., multiplied by the mean of feature importances), S_A_FI (the mean of sign alignments obtained from the Global Alignment Measurement described in subchapter 4.6., multiplied by the mean of feature importances), SR_A_FI (the mean of the vector containing rank and sign alignments obtained from the Global Alignment Measurement described in, multiplied by the mean of feature importances).

In what concerns the comparison between the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), for each feature alignment scheme R, S and SR, we observed that the highest averaged increase was obtained for the R feature alignment scheme (which uses only feature attribution ranks), for the LIME and SHAP explainers. We recorded a positive increase from the averaged (averaged across explanation methods) weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by a single explanation method to the averaged (averaged across feature overlap methods) weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset. This demonstrates that the aggregation strategy helps the explainers to reach a consensus and resolve the disagreement problem more effectively than using a single explanation method.

Pirie et al. (2023) advanced AGREE, an explainer aggregation framework that uses knowledge from the neighborhood spaces of feature attribution vectors. Five explainers (LIME, Kernel SHAP, Deep SHAP, Integrated Gradients and MAPLE) have been used to generate explanations for a set of black-box models trained on 8 regression and classification datasets, including Liver and Abalone, that have been also considered in our research. AGREE beat or matched the performance of the k -NN classifier using averaging weighting of the 6 feature agreement methods proposed by Krishna et al. (2023): Feature Agreement, Sign Agreement, Rank Agreement, Signed Rank Agreement, Rank Correlation, Pairwise Rank Alignment, for 7 out of the 8 datasets. The difference between our research and theirs is that we performed an aggregation across explainers and their multiple explanations (a number of 100 explanations have been generated for each instance from the dataset) and that we hybridized the feature overlap explanation weight methods by integrating and influential classification factor – the importance (information gain) calculated by each feature.

7 Conclusions

We presented a global cluster-based aggregation framework inspired from textual Case Based Reasoning for rectifying the Rashômon Effect - the disagreement between the explanations provided by various attribution explainers and the dissimilarity between multiple explanations generated by a single explainer for a particular instance from the dataset. The implemented Global Alignment Method aligns multiple explanations provided by 3 popular feature attribution explainers (LIME, SHAP and Anchors) by comparing problem and solution space clusters and applying the "case base image" metaphor that retrieves patterns, regularities and associations between the explanations. We proposed 9 feature overlap explanation schemes for explainers and explanations alignment, that used either the mean of ranked feature attribution vectors, the mean of feature importances, the mean of rank alignments, the mean of sign alignments, the mean of the vector containing rank and sign alignments (all three obtained from the Global Alignment Measurement) and a hybridized vector of ranked feature attribution vectors / rank alignments / sign alignments / vector of sign and rank alignments, that integrated the importance of each feature, for enhanced classification robustness and evaluation performance. The hybridization process refers to the multiplication with the mean of feature importances. The resulting aggregated explanation weight vectors have been provided to the feature space of a weighted k -NN algorithm and we compared the prediction performance against a non-weighted k -NN predictor, having as task the binary classification for 6 state-of-the-art datasets. The proposed aggregation strategy based on the Global Alignment Measurement for rank alignment, hybridized with feature importance scores, performed equally as well as the average of ranked feature attributions, also hybridized with feature importance scores. They showed significant performance improvements for all three evaluation metrics (accuracy, F1-score and ROC-AUC score) compared to the non-weighted k -NN for 5 out of 6 datasets and outperformed the rest of feature overlap explanation methods. They were closely followed by the feature overlap explanation methods that employed sign alignments and both sign and rank alignments, both methods hybridized with feature importance scores. We therefore highlight the significance of feature importance when applying and evaluating the feature overlap explanation weight methods, the role they have in measuring explainers' disagreement and explanations dissimilarity, as well as the support they offer in effective model learning. The results are encouraging, as they suggest that the intuition of merging feature importances with global averages or global alignments across explainers and their multiple explanations generated iteratively is more successful than local alignment or the simple mean of feature rankings proposed by previous studies (Pirie et al. 2023). This can increase the confidence in XAI models, which is very important to their embracement in real-world usage.

In what concerns the comparison between the results of the weighted k -NN algorithm using aggregated feature overlap explanation weights to the weighted k -NN algorithm using weights produced by a single explanation method (either LIME, SHAP or Anchors), for each feature alignment scheme R, S and SR, we observed that the aggregation strategy helps the explainers to reach a consensus and resolve the disagreement problem more effectively than using a single explanation method.

As future research directions, we plan to extend our study by evaluating the method on regression tasks, by enlarging the number of tested datasets and by verifying if the results are similar on additional explanation methods, such as gradient-based algorithms.

Appendix

Method evaluation

Aggregated feature weighted classification performance vs nonweightedclassification performance Tables 4, 5, 6.

Comparison between the averaged metrics of the weighted k-NNpredictions using each of the 9 feature overlap explanation weightsto the predictions of the non-weighted k-NN classifier.

For example, for the Diabetes dataset, we obtained that only the averaged accuracy scores of the weighted k -NN classifier using the S_A feature overlap explanation weights were not different from the averaged accuracy scores generated by the non-weighted k -NN classifier (did not reject the null hypothesis). The same situation happened in the case of the Liver dataset for the R_FI weights, and so on.

We then computed the non-parametric Friedman Chi Square test to verify if the averaged performance metrics of the repeated measured data were different across groups. For the accuracy metric, the p -value of the Friedman test was 0.00019, for the F1-score, p -value=6.69e-05 and for the ROC-AUC score, p -value=0.00022. The p -values were less than the significance threshold of 0.05, therefore we concluded that the averaged accuracies, F1-scores and ROC-AUC scores of the weighted k -NN and non-weighted k -NN predictions were significantly different (Tables 7 and 8).

Comparison between the average increase (in %) of a feature overlap explanation weight method performance to the rest of the feature overlap explanation weight methods.

Tables 9, 10 and 11 present the datasets for which the Holm-Bonferroni corrected Wilcoxon signed-rank test determined that the pair of feature overlap explanation weight methods corresponding to row and column are significantly different. The datasets first letters are colored either in green or red. For those colored in green, the feature overlap explanation weight method indicated on the row, when fitted to the weighted k -NN algorithm, conducts to a higher classification performance (accuracy for Table 9, F1-score for Table 10 and ROC-AUC score for Table 11) than the feature explanation weight method specified on the column for the weighted k -NN algorithm (Table 12).

Table 4 Accuracy scores of the weighted k -NN using each of the 9 feature overlap explanation weights vs. non-weighted k -NN

	Accuracy (%)										
Dataset	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average	Non-weighted k -NN
Diabetes	74.97 (p=0.0001)	74.13 (p=0.01)	74.7 (p=0.0006)	74.17 (p=0.009)	**75.13** (p=1.6e-05)	(p=0.059>0.05)	73.5 (p=0.0009)	73.77 (p=0.03)	74 (p=0.01)	74.3	72.58
Liver	68.59 (p=3.1e-05)	(p=0.41>0.05)	**69.04** (p=4.02e-06)	67.90 (p=0.0004)	68.66 (p=1.29e-05)	67.26 (p=0.01)	67.3 (p=0.01)	67.12 (p=0.01)	67.57 (p=0.009)	67.93	65.65
Hepatitis	(p=0.16>0.05)	95.49 (p=2.59e-09)	**95.54** (p=5.7e-12)	(p=0.1>0.05)	95.50 (p=3.55e-15)	93.3 (p=0.01)	94.43 (p=1.58e-05)	(p=0.08>0.05)	94.84 (p=2.86e-08)	94.85	92.43
Abalone	(p=0.29>0.05)	**54.47** (p=0.0002)	52.67 (p=0.004)	(p=0.06>0.05)	(p=0.02, but rejected by the Bonferroni correction)	(p=0.56>0.05)	(p=0.09>0.05)	(p=0.11>0.05)	(p=0.47>0.05)	53.55	53.62
Water	61.40 (p=2.15e-05)	62.33 (p=1.79e-06)	62.14 (p=6.35e-05)	61.86 (p=0.001)	**62.37** (p=8.88e-07)	61.47 (p=0.008)	62.23 (p=7.57e-06)	(p=0.76>0.05)	61.98 (p=0.0002)	61.97	60.86
Fetal	92.51 (p=0.0001)	92.82 (p=1.78e-09)	93.47 (p=3.51e-09)	92.53 (p=0.0006)	**93.52** (p=2.63e-12)	92.2 (p=0.01)	92.94 (p=1.66e-06)	(p=0.09, but rejected by the Bonferroni correction)	92.96 (p=8.76e-07)	92.87	91.74

The maximum accuracy for each dataset values are highlighted in bold

Table 5 F1 scores of the weighted k -NN using each of the 9 feature overlap explanation weights vs. non-weighted k -NN

	F1 score (%)										
Dataset	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average	Non-weighted k -NN
Diabetes	61.95 (p=8.4e-08)	59.86 (p=9.46e-05)	62.18 (p=1.9e-08)	61.33 (p=8.47e-08)	**62.79** (p=8.85e-10)	(p=0.09>0.05)	57.95 (p=4.07e-05)	(p=0.1>0.05)	58.64 (p=0.0006)	60.67	55.74
Liver	78.46 (p=3.45e-05)	(p=0.95>0.05)	**78.57** (p=0.0001)	77.84 (p=0.002)	78.42 (p=4.87e-05)	77.98 (p=0.001)	77.87 (p=0.005)	77.59 (p=0.01)	77.93 (p=0.006)	78.08	76.45
Hepatitis	(p=0.3>0.05)	77.94 (p=1.77e-15)	**79.07** (p=5.32e-15)	62.17 (p=0.003)	78.6 (p=1.77e-15)	63.65 (p=0.0002)	74.52 (p=4.49e-13)	61.38 (p=0.001)	74.7 (p=1.77e-14)	71.51	55.01
Abalone	(p=0.39>0.05)	**46.44** (p=0.001)	(p=0.01, but rejected by the Bonferroni correction)	(p=0.01, but rejected by the Bonferroni correction)	(p=0.04, but rejected by the Bonferroni correction)	(p=0.46)	(p=0.33)	(p=0.88)	(p=0.45)	46.44	47.78
Water	43.72 (p=3.38e-06)	44.56 (p=0.0005)	44.4 (p=0.0001)	44.2 (p=0.002)	**45.1** (p=1.13e-06)	(p=0.06>0.05)	44.5 (p=0.0001)	(p=0.85>0.05)	44.8 (p=1.74e-06)	72.35	42.81
Fetal	82.15 (p=9.35e-06)	82.91 (p=1.77e-14)	84.79 (p=3.67e-13)	82 (p=0.0004)	**84.82** (p=4.44e-14)	81.11 (p=0.01)	83.41 (p=8e-10)	(p=0.03, rejected by the Bonferroni correction)	83.17 (p=2.43e-08)	83.04	79.95

The maximum F1 score for each dataset values are highlighted in bold

Table 6 ROC-AUC scores of the weighted k-NN using each of the 9 feature overlap explanation weights vs. non-weighted k -NN

	ROC AUC score (%)										
Dataset	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average	Non-weighted k -NN
Diabetes	71.28 (p=9.74e-08)	69.92 (p=0.0002)	71.35 (p=4.26e-07)	70.69 (p=3.99e-07)	**71.93** (p=2.26e-10)	(p=0.084>0.05)	68.66 (p=0.0001)	(p=0.066>0.05)	69.15 (p=0.001)	70.42	67.34
Liver	59.88 (p=5.68e-06)	(p=0.11>0.05)	**61.4** (p=8.22e-09)	59.63 (p=2.68e-05)	60.39 (p=4.26e-07)	(p=0.81>0.05)	(p=0.18>0.05)	(p=0.067>0.05)	(p=0.055>0.05)	60.33	56.41
Hepatitis	73.45 (p=0.001)	83.82 (p=1.77e-15)	**85.02** (p=1.77e-15)	(p=0.003>0.05)	84.54 (p=1.77e-15)	73.89 (p=0.0004)	82.83 (p=1.77e-14)	72.91 (p=0.001)	81.81 (p=3.55e-15)	79.78	69.59
Abalone	(p=0.3>0.05)	**52.07** (p=0.0001)	(p=0.009, but rejected by the Bonferroni correction)	(p=0.003, but rejected by the Bonferroni correction)	(p=0.003, but rejected by the Bonferroni correction)	(p=0.81>0.05)	(p=0.08>0.05)	(p=0.12>0.05)	(p=0.57>0.05)	52.07	53.13
Water	57.28 (p=1.06e-06)	58.20 (p=8.23e-07)	57.92 (p=7.51e-06)	57.70 (p=0.0008)	**58.28** (p=9.74e-08)	57.293 (p=0.008)	58.04 (p=3.80e-06)	(p=0.98>0.05)	58 (p=3.38e-06)	57.84	56.68
Fetal	87.17 (p=0.0002)	88.09 (p=8.8852e-15)	89.01 (p=3.88e-10)	87.05 (p=0.004)	**89.35** (p=8.88e-15)	(p=0.20>0.05)	88.19 (p=3.095e-08)	(p=0.26>0.05)	88.06 (p=4.26e-07)	88.13	85.88

The maximum ROC-AUC score for each dataset values are highlighted in bold

Table 7 Increase (in %) from non-weighted k-NN compared to averaged weighted k -NN

Dataset	Increase (in %) (non-weighted k -NN compared to averaged weighted k -NN)		
	Accuracy	F1-score	ROC-AUC score
Diabetes	1.71	4.93	3.08
Liver	2.28	1.63	3.92
Hepatitis	2.41	16.49	10.19
Abalone	-0.07	-1.34	-1.06
Water	1.10	1.65	1.15
Fetal	1.12	3.09	2.24

Table 8 Increase (in %) from the best performing feature overlap weighted k -NN model compared to the non-weighted k -NN

Dataset	Increase (in %) (best performing feature overlap weighted k-NN model compared to the non-weighted k-NN)		
	Accuracy	F1-score	ROC-AUC score
Diabetes (R_A_FI)	2.54	7.05	4.59
Liver (R_AVG_FI)	3.39	2.12	4.98
Hepatitis (R_AVG_FI)	3.10	24.05	15.43
Abalone (R_FI)	0.84	-1.34	-1.06
Water (R_A_FI)	1.5	2.27	1.59
Fetal (R_A_FI)	1.78	4.86	3.46

Legend for Tables 9, 10 and 11:
- D – Diabetes dataset
- L – Liver dataset
- H – Hepatitis Dataset
- A – Abalone Dataset
- W – Water dataset
- F – Fetal dataset

Table 13 presents the accuracy of the k -NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the accuracy of k -NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the accuracy of R_AVG_FI. The average column contains the mean of the accuracies of the k -NN predictor when using the 8 feature overlap explanation methods (except for R_AVG_FI) and the increase (%) column presents the difference between the accuracy obtained when using R_AVG_FI and the average column, for each dataset.

Table 9 Datasets for which the accuracy scores of the weighted k -NN algorithm are significantly different when using the feature overlap explanation weights computed based on the methods specified on the rows and on the columns

	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI
R_AVG		LHAW	HWF		HWF		HWF	D	H
R_FI	LHAW		LAF	LHA	LAF	HAWF		HWF	A
R_AVG_FI	HWF	LAF		HF		LHAF	AF	LHAWF	AF
R_A		LHA	HF		HF	A	HA	AW	HA
R_A_FI	HWF	LAF		HF		HWF	DAF	DHAWF	AF
S_A		HWF	LHAWF	A	HWF		HWF		HF
S_A_FI	HWF		AF	HA	AF	HWF		HWF	
SR_A	D	HWF	LHAWF	AW	DHAWF		HWF		HWF
SR_A_FI	H	A	AF	HA	AF	HF		HWF	

Table 14 presents the accuracy of the k -NN classifier when using the R_A_FI feature overlap weights (in bold) and also the accuracy of k -NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the accuracy of R_A_FI. The average column contains the mean of the accuracies of the k -NN predictor when using the 8 feature overlap explanation methods (except for R_A_FI) and the increase (%) column presents the difference between the accuracy obtained when using R_A_FI and the average column, for each dataset.

Table 10 Datasets for which the F1 scores of the weighted k -NN algorithm are significantly different when using the feature overlap explanation weights computed based on the methods specified on the rows and on the columns

	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI
R_AVG		L H A	H F		H W F	D	D H F	D	D H W
R_FI	L H A		D L A F	L H A	D L A F	D L H A W F	L H A	D L H A W F	L H A
R_AVG_FI	H F	D L A F		H F		D H F	D H A F	D H W F	D H F
R_A		L H A	H F		H F	D	D H A F	D	D H
R_A_FI	H W F	D L A F		H F		D H W F	D A F	D H W F	D H F
S_A	D	D L H A W F	D H F	D	D H W F		H F		H W F
S_A_FI	D H F	A	D H A F	D H A F	D A F	H F		H W F	
SR_A	D	D H A W F	D H W F	D	D H W F		H W F		H W F
SR_A_FI	D H W	H A	D H F	D H	D H F	H W F		H W F	

Table 15 presents the F1-score of the k -NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the F1-score of k -NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the F1-score of R_AVG_FI. The average column contains the mean of the F1-scores of the k -NN predictor when using the 8 feature overlap explanation methods (except for R_AVG_FI) and the increase (%) column presents the difference between the F1-score obtained when using R_AVG_FI and the average column, for each dataset.

Table 11 Datasets for which the ROC AUC scores of the weighted k -NN algorithm are significantly different when using the feature overlap explanation weights computed based on the methods specified on the rows and on the columns

	R_AVG	R_FI	R_AVG_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI
R_AVG		L H A W	H F		H W F	D L	D L H W F	D L W	D H W
R_FI	L H A W		L A	L H A	D L A F	H A W F	A	H A W F	A
R_AVG_FI	L H F	L A F		L H F		D L H A F	D L H A	D L H A F	D H F
R_A		L H A W	L H F		D H F	D L A	D L H A F	D L A W	D H A
R_A_FI	H W F	D L A F		D H F		D L H W F	D L A F	D L H A W F	D L H F
S_A	D L	D H A W F	D L H A F	D L A	D L H W F		H F		H W F
S_A_FI	D L H W F		D L A	D L H A F	D L A F	H F		H W F	
SR_A	D L W	H W F	D L H A W F	D L A W	D L H A W F		H W F		H W F
SR_A_FI	D H W	H A	D L H F	D H A	D L H F	H W F		H W F	

Table 12 Average increase (in %) of a feature overlap explanation weight method performance (accuracy, F1-score and ROC-AUC score), compared to the rest of the feature overlap explanation weight methods, for all datasets

Dataset	Average increase (in %) of a feature overlap explanation weight method performance compared to the rest		
	Accuracy	F1-score	ROC-AUC score
R_AVG	-0.67	-1.72	-1.07
R_FI	-0.08	1.57	0.86
R_AVG_FI	1.57	4.03	3.56
R_A	0.2	-2.05	-1.35
R_A_FI	1.63	4.34	3.2
S_A	-1.26	-3.08	-3.17
S_A_FI	0.33	1.24	0.66
SR_A	-1.45	-4.63	-3.25
SR_A_FI	0.65	0.82	-0.07

Table 16 presents the F1-score of the k-NN classifier when using the R_A_FI feature overlap weights (in bold) and also the F1-score of k-NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the F1-score of R_A_FI. The average column contains the mean of the F1-scores of the k-NN predictor when using the 8 feature overlap explanation methods (except for R_A_FI) and the increase (%) column presents the difference between the F1-score obtained when using R_A_FI and the average column, for each dataset.

Table 17 presents the ROC-AUC score of the k-NN classifier when using the R_AVG_FI feature overlap weights (in bold) and also the ROC-AUC score of k-NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the ROC-AUC score of R_AVG_FI. The average column contains the mean of the ROC-AUC scores of the k-NN predictor when using the 8 feature overlap explanation methods (except for R_AVG_FI) and the increase (%) column presents the difference between the ROC-AUC score obtained when using R_AVG_FI and the average column, for each dataset.

Table 18 presents the ROC-AUC score of the k-NN classifier when using the R_A_FI feature overlap weights (in bold) and also the ROC-AUC score of k-NN when employing the other 8 feature overlap weights, for each dataset. The empty cells signify that the respective metric performances were not statistically different from the ROC-AUC score of R_A_FI. The average column contains the mean of the ROC-AUC scores of the k-NN predictor when using the 8 feature overlap explanation methods (except for R_A_FI) and the increase (%) column presents the difference between the ROC-AUC score obtained when using R_A_FI and the average column, for each dataset.

Aggregated feature weighted classification performance vs singleexplanation feature weighted classification performance (Tables 19, 20, 21, 22, 23, 24, 25, 26).

Table 13 Comparison to the R_AVG_FI feature overlap explanation method for the accuracymetric

Dataset	Accuracy (%)										
	R_AVG_FI	R_AVG	R_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_AVG_FI)	Increase (%)
Diabetes	**74.7**										
Liver	**69**		65.9			67.3		67.1		**66.75**	**2.25**
Hepatitis	**95.5**	93		93		93.3		93.0		**93.07**	**2.43**
Abalone	**54.5**		52.6			53.5	53.2	53.3	53.5	**53.21**	**1.29**
Water	**62.1**	61.4						60.9		**61.13**	**0.97**
Fetal	**93.5**	92.5	92.8	92.5		92.2	92.9	92.1	93.0	**92.59**	**0.91**

The accuracy of the R_AVG_FI feature overlap explanation method, the average and the increase in percentage values are highlighted in bold

Table 14 Comparison to the R_A_FI feature overlap explanation method for the accuracymetric

Dataset	Accuracy (%)										
	R_A_FI	R_AVG	R_FI	R_AVG_FI	R_A	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_A_FI)	Increase (%)
Diabetes	**75.1**						73.5	73.8		**73.64**	**1.46**
Liver	**68.7**		65.9							**65.86**	**2.84**
Hepatitis	**95.5**	93.0			93.0	93.3		93.0		**93.07**	**2.43**
Abalone	**54.2**		52.6				53.2	53.3	53.5	**53.15**	**1.05**
Water	**62.3**	61.4				61.5		60.9		**61.25**	**1.05**
Fetal	**93.5**	92.5	92.8		92.5	92.2	92.9	92.1	93.0	**92.59**	**0.91**

The accuracy of the R_A_FI feature overlap explanation method, the average and the increase in percentage, for each dataset values are highlighted in bold

Table 15 Comparison to the R_AVG_FI feature overlap explanation method for the F1-scoremetric

Dataset	F1 – score (%)										
	R_AVG_FI	R_AVG	R_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_AVG_FI)	Increase (%)
Diabetes	**62.2**		59.9			57.3	57.9	56.9	58.6	**58.14**	**4.06**
Liver	**78.6**		76.5							**76.51**	**2.09**
Hepatitis	**79.1**	62.4		62.2		63.7	74.5	61.4	74.7	**66.47**	**12.63**
Abalone	**48.6**		46.4				47.5			**46.96**	**1.64**
Water	**44.4**							43.1		**43.08**	**1.32**
Fetal	**84.8**	82.1	82.9	82		81.1	83.4	81.1	83.2	**82.26**	**2.54**

The F1 score of the R_AVG_FI feature overlap explanation method, the average and the increase in percentage, for each dataset values are highlighted in bold

Table 16 Comparison to the R_A_FI feature overlap explanation method for the F1-scoremetric

Dataset	F1 – score (%)										
	R_A_FI	R_AVG	R_FI	R_AVG_FI	R_A	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_A_FI)	Increase (%)
Diabetes	**62.8**		59.9			57.3	57.9	56.9	58.6	**58.14**	**4.66**
Liver	**78.4**		76.5							**76.51**	**1.89**
Hepatitis	**78.6**	62.4			62.2	63.7		61.4	74.7	**64.86**	**13.74**
Abalone	**48.5**		46.4				47.5			**46.96**	**1.54**
Water	**45.1**	43.7				43.5		43.1		**43.44**	**1.66**
Fetal	**84.8**	82.1	82.9		82	81.1	83.4	81.1	83.2	**82.26**	**2.54**

The F1 score of the R_A_FI feature overlap explanation method, the average and the increase in percentage, for each dataset values are highlighted in bold

Table 17 Comparison to the R_AVG_FI feature overlap explanation method for the ROC-AUCscore metric

Dataset	ROC-AUC score (%)										
	R_AVG_FI	R_AVG	R_FI	R_A	R_A_FI	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_AVG_FI)	Increase (%)
Diabetes	**71.3**					68.4	68.6	68.3	69.1	**68.64**	**2.66**
Liver	**61.4**	59.8	56.9	59.6		56.8	57.3	57.6		**58.04**	**3.36**
Hepatitis	**85**	73.4		73.2		73.8	82.8	72.9	81.8	**76.35**	**8.65**
Abalone	**53.9**		52			53	52.6	52.8		**52.64**	**1.26**
Water	**57.9**										
Fetal	**89**	87.1	88	87		86.4		86.3	88	**87.19**	**1.81**

The ROC-AUC score of the R_AVG_FI feature overlap explanation method, the average and the increase in percentage, for each dataset values are highlighted in bold

Table 18 Comparison to the R_A_FI feature overlap explanation method for the ROC-AUCscore metric

Dataset	ROC-AUC score (%)										
	R_A_FI	R_AVG	R_FI	R_AVG_FI	R_A	S_A	S_A_FI	SR_A	SR_A_FI	Average (except for R_A_FI)	Increase (%)
Diabetes	**71.9**		69.9		70.6	68.4	68.6	68.3	69.1	**69.19**	**2.71**
Liver	**60.3**		56.9			56.8	57.3	57.6	58	**57.36**	**2.94**
Hepatitis	**84.5**	73.4			73.2	73.8		72.9	81.8	**75.06**	**9.44**
Abalone	**53.6**		52				52.6	52.8		**52.52**	**1.08**
Water	**58.2**	57.2				57.2		58		**57.52**	**0.68**
Fetal	**89.3**	87.1	88		87	86	88.1	86.3	88	**87.33**	**1.97**

The ROC-AUC score of the R_A_FI feature overlap explanation method, the average and the increase in percentage, for each dataset values are highlighted in bold

Table 19 Increase (in %) from the weighted k-NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the LIME explanation method,scheme R, to the averaged (averaged across feature overlap methods) and maximumweighted k-NN performance metric scores where the weights are feature overlapexplanation weights, for each dataset

	LIME *R*					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	ROC-AUC score (%)	Accuracy (%)	F1-score (%)	ROC-AUC score (%)
Diabetes	2.84	7.67	4.85	3.34	8.84	5.74
Liver	0.67	0.30	1.43	1.70	0.91	3.18
Hepatitis	2.12	15.83	10.02	3.15	22.86	15.03
Abalone	0.13	-0.33	0.04	0.73	0.45	0.60
Water	1.78	4.54	2.35	2.13	5.24	2.76
Fetal	1.46	3.67	2.67	2.01	5.16	3.88

Table 20 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the LIME explanation method, scheme R, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	LIME S					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	0.81	2.67	1.48	1.63	4.42	2.41
Liver	0.27	0.00	0.75	0.51	0.55	0.17
Hepatitis	0.84	8.22	5.76	1.83	18.92	13.00
Abalone	-1.00	-1.06	-0.47	-0.65	-0.29	-0.60
Water	1.60	4.00	2.12	2.93	4.38	3.01
Fetal	1.27	3.11	2.19	1.47	4.03	2.90

Table 21 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the LIME explanation method, scheme S, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	LIME SR					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	1.72	4.24	2.47	1.84	5.11	2.90
Liver	0.55	0.33	0.71	0.78	0.51	0.93
Hepatitis	1.30	12.44	7.53	2.24	19.10	11.98
Abalone	-0.71	-0.32	-0.67	-0.59	-0.16	-0.53
Water	2.11	3.81	2.97	2.68	4.69	2.97
Fetal	1.08	4.24	2.47	1.49	5.11	2.90

Table 22 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the LIME explanation method, scheme SR, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	SHAP *R*					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	1.68	4.15	2.74	2.18	5.32	3.64
Liver	1.52	0.82	2.63	2.55	1.43	4.38
Hepatitis	2.10	16.54	10.19	3.12	23.58	15.20
Abalone	1.20	1.48	1.23	1.80	2.27	1.79
Water	2.07	3.15	2.21	2.42	3.85	2.61
Fetal	1.68	1.07	0.85	2.18	2.55	2.07

Table 23 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the SHAP explanation method, scheme R, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	SHAP S					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	0.77	1.33	0.80	0.64	0.48	0.30
Liver	0.16	0.08	0.09	1.05	0.92	0.89
Hepatitis	1.73	14.36	8.76	2.00	18.09	12.48
Abalone	0.31	0.93	0.37	0.61	1.34	0.72
Water	2.23	3.17	2.27	3.04	3.53	2.88
Fetal	0.11	0.39	0.43	0.54	1.64	1.37

Table 24 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the SHAP explanation method, scheme S, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	SHAP SR					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	0.74	0.31	0.36	0.85	1.18	0.79
Liver	1.09	0.70	1.43	1.31	0.87	1.65
Hepatitis	1.47	11.61	7.01	2.41	18.27	11.45
Abalone	0.54	1.32	0.65	0.67	1.47	0.79
Water	2.23	2.96	2.84	2.79	3.84	2.84
Fetal	0.15	0.31	0.36	0.56	1.18	0.79

Table 25 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the SHAP explanation method, scheme SR, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	Anchors R					
	Average			Maximum		
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Diabetes	0.33	1.04	0.66	0.84	2.22	1.56
Liver	-0.73	-0.32	-1.46	0.30	0.29	0.29
Hepatitis	-2.52	-14.85	-10.00	-1.49	-7.82	-4.99
Abalone	0.73	1.70	0.84	1.33	2.48	1.40
Water	0.44	0.30	0.37	0.79	1.00	0.77
Fetal	-1.53	-3.68	-2.20	-0.98	-2.20	-0.98

Table 26 Increase (in %) from the weighted k -NN performance metric scores (accuracy, F1-score, ROC-AUC score) where the weights are produced by the anchors explanation method, scheme R, to the averaged (averaged across feature overlap methods) and maximum weighted k -NN performance metric scores where the weights are feature overlap explanation weights, for each dataset

	Increase (in %) (averaged single explanation method weighted $\boldsymbol{k}$ -NN compared to averaged feature aggregated weighted $\boldsymbol{k}$ -NN)		
	Accuracy (%)	F1-score (%)	ROC-AUC score (%)
Diabetes	1.27	3.06	1.91
Liver	0.50	0.27	0.80
Hepatitis	1.01	9.16	5.61
Abalone	0.17	0.53	0.28
Water	1.78	3.13	2.16
Fetal	0.60	1.30	0.97

Acknowledgements This work was supported by project TRAVEE (Virtual Therapist with Augmented Feedback for Neuromotor Recovery) through a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-III-P2-2.1-PTE-2021-0634, within PNCDI III and by a grant of the Petroleum-Gas University of Ploiesti, project number 11061/2023, within Internal Grant for Scientific Research. Marius Leordeanu was supported by EU Horizon project ELIAS (Grant number: 101120237).

Author contributions Conceptualization, O.M. and G.M.; methodology, O.M. and G.M.; software, O.M.; validation, A.M., F.M., M.L.; formal analysis, O.M.; investigation, O.M., G.M., L.P.; resources, A.M.; data curation, M.L.; writing—original draft preparation, O.M.; writing—review and editing, A.M., F.M.; visualization, O.M.; supervision, M.L., F.M., L.P.; project administration, A.M.; funding acquisition, A.M.

Funding This work was supported by project TRAVEE (Virtual Therapist with Augmented Feed-back for Neuromotor Recovery) through a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-III-P2-2.1-PTE-2021-0634, within PNCDI III. Gabriela Moise was supported by a grant of the Petroleum-Gas University of Ploiesti, project number 11061/2023, within Internal Grant for Scientific Research. Marius Leordeanu was supported by EU Horizon project ELIAS (Grant number: 101120237).

Data availability Data Availability Statement: The code and results are publicly available at this address: https://github.com/oanabalan/AggregateExplanations.

Declarations

Competing interests The authors declare no competing interests.

References

Adadi A, Berrada M (2018) Peeking inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aïvodji U, Arai H, Fortineau O, Gambs S, Hara S, Tapp A (2019) Fairwashing: the risk of rationalization. In International Conference on Machine Learning, pp 161–170, 2019

Bayrak B, Bach K (2022) When to Explain? Model Agnostic Explanation Using a Case-based Approach and Counterfactuals. Proceedings of the 34th Norwegian ICT conference for research and education – NIKT 2022 ISBN: 978-3-16-148410-0

Bordt S, Finck M, Raidl E, von Luxburg U (2022) Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, pp. 891–905. https://doi.org/10.1145/3531146.3533153

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231. https://doi.org/10.1007/s10462-023-10420-8

Brownlee J (2020) LOOCV for Evaluating Machine Learning Algorithms. https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/. Accessed January 2024

Brughmans D, Melis L, Martens D (2023) Disagreement amongst counterfactual explanations: How transparency can be deceptive. arXiv [csAI]. http://arxiv.org/abs/2304.12667

Camburu O, Giunchiglia E, Foerster J, Lukasiewicz T, Blunsom P (2019) Can I trust the explainer? verifying posthoc explanatory methods. CoRR, abs/1910.02065, http://arxiv.org/abs/1910.02065

Campos D, Bernardes J (2010) Cardiotocography. UCI Mach Learn Repository. https://doi.org/10.24432/C51S4N. Accessed January 2024

Chakraborti S, Beresi U, Wiratunga N, Massie S, Lothian R, Watt S (2007) A Simple Approach towards Visualizing and Evaluating Complexity of Textual Case Bases. In: Proc. of the ICCBR 2007 Workshops

Chen Y, Hao Y (2017) A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Syst Appl 80:340–355. https://doi.org/10.1016/j.eswa.2017.02.044

Del Giudice M (2021) The prediction-explanation fallacy: a pervasive problem in scientific applications of machine learning. PsyArXiv. https://doi.org/10.31234/osf.io/4vq8f

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. Proc Conf AAAI Artif Intell 33(01):3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

Goethals S, Martens D, Evgeniou T (2023) Manipulation risks in explainable AI: the implications of the disagreement problem. arXiv [csAI]. https://doi.org/10.48550/arXiv.2306.13885

Hastie TJ (2017) Generalized Additive Models. In: Statistical Models in S. Routledge. pp 249–307

Hepatitis (1988) UCI Machine Learning Repository. https://doi.org/10.24432/C5Q59J. Accessed January 2024

Kadiwal A (2021) Water quality dataset. https://www.kaggle.com/datasets/adityakadiwal/water-potability. Accessed January 2024

Krishna S, Han T, Gu A, Jabbari S, Wu ZS, Lakkaraju H (2023) The disagreement problem in explainable machine learning: a practitioner's perspective. Res Square. http://arxiv.org/abs/2202.01602

Kundu RK, Hoque KA (2023) Explainable predictive maintenance is not enough: quantifying trust in remaining useful life estimation. Proc Annu Conf Progn Health Manag Soc 15(1). https://doi.org/10.36001/phmconf.2023.v15i1.3472

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. arXiv [csAI]. http://arxiv.org/abs/1705.07874

Malinka K, Peresíni M, Firc A, Hujnák O, Janus F (2023) On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree? In: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1. New York, NY, USA: ACM

Mikalef P, Gupta M (2021) Artificial intelligence capability: conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Inf Manag 58(3):103434. https://doi.org/10.1016/j.im.2021.103434
Müller S, Toborek V, Beckh K, Jakobs M, Bauckhage C, Welke P (2023) An empirical evaluation of the Rashomon effect in explainable machine learning. Machine learning and knowledge Discovery in databases: Research Track. Springer Nature Switzerland, Cham, pp 462–478
Nash W, Sellers T, Talbot S, Cawthorn A, Ford W (1994) Abalone. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W. Accessed January 2024
Neely M, Schouten SF, Bleeker MJR, Lucic A (2021) Order in the court: Explainable AI methods prone to disagreement. arXiv [csLG]. http://arxiv.org/abs/2105.03287
Pirie C, Wiratunga N, Wijekoon A, Moreno-Garcia CF (2023) AGREE: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. In Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023), pp184–199. CEUR
Poiret C, Grigis A, Thomas J, Noulhiane M (2023) Can we agree? On the Rashomon effect and the reliability of post-hoc explainable AI. arXiv [csLG]. http://arxiv.org/abs/2308.07247
Raghunandan MA, Wiratunga N, Chakraborti S, Massie S, Khemani D (2008) Evaluation measures for TCBR systems. Lecture notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 444–458
Ramana B, Venkateswarlu N (2012) ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. https://doi.org/10.24432/C5D02C. Accessed January 2024
Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM
Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. Proc Conf AAAI Artif Intell 32(1). https://doi.org/10.1609/aaai.v32i1.11491
Roy S, Laberge G, Roy B, Khomh F, Nikanjam A, Mondal S (2022) Why don't XAI techniques agree? Characterizing the disagreements between post-hoc explanations of defect predictions. In: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE
Saarela M, Geogieva L (2022) Robustness, Stability, and Fidelity of explanations for a deep skin Cancer classification model. Appl Sci 12(19):9545. https://doi.org/10.3390/app12199545
Selvaraju RR, Cogswell M, Das A et al (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128:336–359. https://doi.org/10.1007/s11263-019-01228-7
Shamsabadi AS, Yaghini M, Dullerud N, Wyllie S, Aïvodji U, Alaagib A, Gambs S, Papernot N (2022) Washing The Unwashable: On The (Im)possibility of Fairwashing Detection Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022). https://proceedings.neurips.cc/paper_files/paper/2022/hash/5b84864ff8474fd742c66f219b2eaac1-Abstract-Conference.html
Shapley LS (1953) A value for n-Person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games (AM-28), volume II. Princeton University Press, Princeton, pp 307–318
Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) How can we fool LIME and SHAP? Adversarial attacks on post hoc explanation methods. In AAAI/ACM Conference on AI, Ethics, and Society, pp180–186
Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) SmoothGrad: removing noise by adding noise. arXiv [csLG]. http://arxiv.org/abs/1706.03825
Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care, pp 261–265. IEEE Computer Society Press
Stahl BC, Antoniou J, Bhalla N, Brooks L, Jansen P, Lindqvist B, Kirichenko A, Marchal S, Rodrigues R, Santiago N et al (2023) A systematic review of artificial intelligence impact assessments. Artif Intell Rev 1–33. https://doi.org/10.1007/s10462-023-10420-8
Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. arXiv [csLG]. http://arxiv.org/abs/1703.01365
Velmurugan M, Ouyang C, Moreira C, Sindhgatta R (2021) Developing a fidelity evaluation approach for interpretable machine learning. arXiv [csLG]. http://arxiv.org/abs/2106.08492
Wolff J, Pauling J, Keck A, Baumbach J (2020) The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 22(2):e16866. https://doi.org/10.2196/16866
Yalcin O, Fan X, Liu S (2021) Evaluating the correctness of explainable AI algorithms for classification. arXiv [csAI]. http://arxiv.org/abs/2105.09740

Authors and Affiliations

Oana Mitruț[1] · Gabriela Moise[2] · Alin Moldoveanu[1] · Florica Moldoveanu[1] · Marius Leordeanu[1] · Livia Petrescu[3]

✉ Alin Moldoveanu
alin.moldoveanu@upb.ro

Oana Mitruț
oana.balan@upb.ro

Gabriela Moise
gmoise@upg-ploiesti.ro

Florica Moldoveanu
florica.moldoveanu@upb.ro

Marius Leordeanu
marius.leordeanu@upb.ro

Livia Petrescu
livia.petrescu@bio.unibuc.ro

[1] Faculty of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest, 060042 Bucharest, Romania

[2] Faculty of Letters and Sciences, Petroleum-Gas University of Ploiesti, 100680 Ploiesti, Romania

[3] Faculty of Biology, University of Bucharest, 010014 Bucharest, Romania

5-15-2019

AUTOMATIC ADAPTATION OF EXPOSURE INTENSITY IN VR ACROPHOBIA THERAPY, BASED ON DEEP NEURAL NETWORKS

Oana Balan
University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, oana.balan@cs.pub.ro

Gabriela Moise
University Petroleum-Gas, gpmoise@gmail.com

Alin Moldoveanu
University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, alin.moldoveanu@cs.pub.ro

Florica Moldoveanu
University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, florica.moldoveanu@cs.pub.ro

Marius Leordeanu
University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, marius.leordeanu@cs.pub.ro

AUTOMATIC ADAPTATION OF EXPOSURE INTENSITY IN VR ACROPHOBIA THERAPY, BASED ON DEEP NEURAL NETWORKS

Research paper

Oana Bălan, University POLITEHNICA of Bucharest, oana.balan@cs.pub.ro

Gabriela Moise, Petroleum-Gas University of Ploiești, gmoise@upg-ploiesti.ro

Alin Moldoveanu, University POLITEHNICA of Bucharest, alin.moldoveanu@cs.pub.ro

Marius Leordeanu, University POLITEHNICA of Bucharest, marius.leordeanu@cs.pub.ro

Florica Moldoveanu, University POLITEHNICA Bucharest, florica.moldoveanu@cs.pub.ro

Abstract

This paper proposes a real-time Virtual Reality game for treating acrophobia that automatically tailors in-game exposure to heights to the players' individual characteristics – affective state and physiological features. The elements of novelty are the automatic estimation of fear and the prediction of the next game level based on the electroencephalogram (EEG) and biophysical data – Galvanic Skin Response (GSR) and Heart Rate (HR). Two neural networks have been trained with the data recorded in an experiment where 4 subjects have been in-vivo and virtually exposed to various heights. In order to test the validity of the approach, the same users played the acrophobia game, using two modalities of expressing fear level. After completing a game level, the EEG and biophysical data were averaged and one neural network estimated the current fear score, while the other predicted the next game level. A measure of similarity between the self-estimated fear level during a game epoch and the fear level predicted by the first neural network showed an accuracy rate of 73% and 42% respectively for the two modalities of expressing fear level. 3 out of 4 users succeeded to obtain a fear level of 0 (complete relaxation) in the final game epoch.

Keywords: Virtual Reality, Gamification, Deep learning, Acrophobia, Emotion Recognition

1 Introduction

Even if the mental health disorders are widely under-reported, there some studies that indicate their worldwide high incidence. According to (Ritchie and Roser, Institute of Health Metrics and Evaluation, World Health Organization (WHO) Global Health Observatory (GHO)), in 2016, an estimation of the number of people with any mental and substance use disorders was about 1.1 billion, from which 275 million were affected by anxiety disorders. The prevalence per country varies between 2.5% and 6.5% and the highest incidence was found in the US, Canada, West and North Europe, Australia, North Africa and South America. Women are more affected than men, with about 4.5% (Ritchie and Roser, 2019). The number of people living with anxiety disorders increased in 2015 with 14.9% since 2005, the estimation being around 264 million persons (WHO, 2017).

Specific phobia is a type of anxiety or fear-related disorder, as classified in the ICD-11 for Mortality and Morbidity Statistics (ICD-11-MMS). Specific phobia is defined *as a marked and excessive fear or anxiety that consistently occurs when exposed to one or more specific objects or situations (e.g., proximity to certain animals, flying, heights, closed spaces, sight of blood or injury) and that is out of proportion to actual danger* (ICD, 2018). In terms of statistics related to phobias, there are studies which estimate that 15-20% of the world's population experience specific phobias at least once in the lifetime (Olesen, 2015). The most common phobias concern heights and animals (Eaton, 2018). The following results have been obtained in a study which involved 22 countries between 2001 and 2011: *the cross-national lifetime and 12-month prevalence rates of specific phobia were, respectively, 7.4% and 5.5%, being higher in females (9.8% and 7.7%) than in males (4.9% and 3.3%) and higher in high and higher-middle income countries than in low/lower middle income countries* (Wardenaar et al, 2017).

Phobias are generally treated with medication and/or psychotherapy. A successful type of psychotherapy is *Cognitive-Behavioural Therapy* (CBT) - with two methods: *cognitive* and *exposure (behavioural) therapy*. Exposure therapy consists in gradual exposure to anxiety eliciting objects or situations, in the presence of a therapist. Virtual Reality is an emergent technology which begins to be adopted more often in phobias therapy. It simulates worlds full of anxiety-producing stimuli and exposes the patients to them in a safe and controlled manner.

In this paper we propose *a Virtual Reality (VR) game for treating acrophobia, using a real-time adaptation of in-game height exposure*. 7.5% of the world's population suffers from acrophobia. 10% of the U.S. population and 14% of the people from U.K. are afraid of heights. Thus, we aim to find a solution based on VR for treating this prevalent anxiety disorder. Using physiological signals (heart rate and galvanic skin response originating from the peripheral nervous system and EEG, stemming from the central nervous system), we feed two Deep Neural Networks (DNNs) in order *to estimate the subject's current fear level and to predict the game level to be played next.* In order to validate our method, we performed an experiment with 4 acrophobic users and observed a high correspondence between the fear level predicted by the neural network and the self-estimated subjective anxiety scores. In addition, 3 out of 4 users succeeded to relax and obtained a fear level of 0 in the last game epochs, concluding that the game levels have been adjusted according to the subjects' emotional state, a fundamental aspect in acrophobia treatment. The presented experiment is part of a series of experiments taking place within a project whose goals is developing a VR system for treating various phobias.

The paper is organized as follows: chapter 2 presents the most relevant VR-based systems for phobia therapy, chapter 3 introduces the relationship between emotions and biophysical data, chapter 4 details the machine learning techniques used for emotion recognition, chapter 5 describes our approach for heights exposure adaptation based on deep neural networks, chapter 6 presents the game design, experimental procedure and research results while finally, chapter 7 shows the conclusions and discusses future work directions.

2 Virtual Reality for Phobia Therapy

Behavioural therapy may be difficult for patients who have problems in imagining scenes full of anxiety producing stimuli and/or who are too afraid to be exposed in real situations. The first use of VR techniques in exposure therapy was reported by the Human-Computer Interaction Group at Clark Atlanta University in November 1992 (North et al, 1997). The first pilot experiments in using *Virtual Reality Exposure Therapy (VRET)* were conducted by North's team for specific phobia treatment: fear of heights, flying, public speaking and fear of being in certain situations (North et al, 1997). Since then, more studies have been undertaken and the results showed that VRET is highly effective and preferred by the patients. In 2 experiments described in (Garcia-Palacios et al, 2001) more than 80% of the subjects (81% and 89%, respectively) chose VRET instead of in-vivo exposure therapy. Also, VRET offered valuable results in the post-treatment assessments, comparable with traditional behavioural therapy (Opris et al, 2012).

One of the largest experiments for acrophobia treatment using VR was performed between October 2017 and February 2018 by a team led by Prof. Freeman from University of Oxford, Department of Psychiatry (Freeman et al, 2018). Using a software application called *Now I Can Do Heights*, the team proved that immersive VR technologies are highly effective for reducing of fear of heights. The procedure did not involve the presence of a therapist, as he was replaced by a virtual coach (Freeman et al, 2018). More examples of VRET are provided by Levski: Bravemind (University of Southern California, Institute for Creative Technologies), VR-based therapeutic solutions for hospital patients developed by Cedars-Sinai, VR therapy for patients with fear of heights, elevators, thunderstorms, flying and public speaking offered by Duke Psychiatry and Behavioural Sciences, Richie's Plank Experience, CityScapes and Landscapes offered by Samsung, Limelight developed by Virtual Neuroscience Lab, Gide Meditation VR developed by Cubicle Ninjas, Relax Soothe Sleep: The Nap App created by Virtually Better, Inc., Limbix VR, Psious.

Gamification elements have been integrated in VRET. *The Climb*, designed for the Oculus device uses as input the Xbox gamepad, the Rift's head tracking and motion tracking, while *Richie's Plank Experience* for HTC Vive (HTC Vive) uses a customizable real plank replicated in the virtual environment (Robertson, 2016).

C2Phobia (C2Phobia) gradually exposes the users to different heights, from the first to the 15th floor of a skyscraper. *Stim Response Virtual Reality* offers a wide range of VR worlds, which are changed using the players' biophysical responses and VR events (2BIOPAC). *Acrophobia Therapy with Virtual Reality (AcTiVity-System)* (Schafer, 2015) – uses an Oculus Rift device to render the 3D scenes. The participants who played the game using an avatar related positively to the approach of the game and even tried to control the system through physical interaction with their bodies. The Stim Response Virtual Reality system (2BIOPAC) offers of wide range of VR worlds. The events from VR and physiological data are synchronized in real-time, while the scenes are changed based on the player's biophysical responses. The Virtual Reality Medical Center (VRMC) used simulation technologies for anxiety and phobia alleviation and educational purposes. VRMC treats patients suffering from panic attacks, specific phobias such as agoraphobia, social phobia, claustrophobia, arachnophobia, fear of flying, fear of driving, fear of thunderstorms, fear of public speaking, using Virtual Reality-enhanced Cognitive Behavioural Therapy (VR-CBT).

In this paper, we continue our previous work (Bălan et al, 2018, Bălan et al, 2019) and propose a VRET system for treating acrophobia, in which patients' data - HR, GSR and EEG are used for fear evaluation and automatic change of VR scenarios. Deep Neural Networks (DNNs) are used for fear classification and automatic height exposure estimation. *As far as we know, there is no VRET system for treating acrophobia based on physiological data and machine learning techniques*.

3 The Relationship between Emotions and Biophysical Data

Emotions are classified using the Circumplex Model of Affects proposed by Russell (Russell, 1979), which consists of two orthogonal emotion dimensions, namely *arousal* and *valence*. Arousal ranges from "not excited" to "excited", while valence, from "positive" to "negative". A third dimension, dominance, indicates the degree of control the subject possesses over his emotions. Usually, *fear is characterized by low valence, high arousal and low dominance* (Demaree et al, 2005).

The approach-withdrawal model, on the other hand, suggests that the right side of the brain mediates withdrawal-based emotions, while activation in the left cortical area is correlated with approach (or appetitive) mental state changes.

Galvanic Skin Response (GSR) is a reflection of skin conductance / resistance change, measured by electrodes applied on the distal phalanges of the index and middle fingers. GSR is a response of the sympathetic nervous system, along with heart rate. Fear is characterized by an increase of sweat production and, in consequence, of skin conductance (DiMeglio, 2015, Healey, 2009, Fleureau et al, 2012, Westerink et al, 2009). Moreover, GSR proved to be efficient in discriminating fear from other negative emotions (AlZoubi et al, 2012). In what concerns heart rate, fear can produce an increase of over 40 bpm from baseline, exceeding the tachycardic threshold of 100-120 bpm (Kometer et al, 2010).

Electroencephalography (EEG) measures brain activity by recording the signals originating from the central nervous system. According to the approach/withdrawal model of frontal alpha asymmetry (Davidson, 1993), left frontal activation, corresponding to low levels of alpha waves (8-12 Hz) indicate a tendency of approach, while, on the other hand, right frontal brain activation (low levels of alpha) elicits negative affective responses (Bos, 2006, Trainor & Schmidt, 2003, Jones & Fox, 1992, Canli et al, 1998). High levels of beta waves (13-30 Hz) indicate anxiety, alert and fear (Arikan et al, 2006, Kometer et al, 2010).

4 Machine Learning for emotion recognition

Emotions play an important role in human communication and interaction. The ability to recognize and differentiate emotions is specific to humans. However, in the last decades, several approaches for automatic identification of emotions have emerged in Emotion Recognition Systems, most of which are using Machine Learning techniques. The most used feature selection algorithms are: Sequential Forward Selection (SFS), Principal Component Analysis (PCA), ANOVA, Fisher's linear discriminant and correlation-based feature selection. The most popular classification techniques are: k-Nearest Neighbours (kNN), Bayesian Networks, Regression Trees, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and artificial networks.

Soleymani et al (Soleymani et al, 2009) used both SVM and a Bayesian framework for classifying different values of arousal and valence into 3 classes – calm, positive excited and negative excited. The Bayesian classifier produced an accuracy of 64% and SVM with linear kernel, which is identical to linear regression, 56%. In (Koelstra et al, 2012), the Fisher linear discriminant was used for feature selection and a Gaussian naïve Bayes classifier for discriminating EEG and peripheral signals into low/high valence, arousal and liking with accuracies of 61%, 64% and 61%. In (Koelstra et al, 2010), the SVM algorithm conducted to an accuracy of 59%/52%/49% using the Power Spectral Density as EEG feature extraction method and 59%/56%/49% with Common Spatial Pattern components extracted from the EEG signals. In the same study, the classification rates for peripheral physiological signals using the SVM classifier and FCBF feature selection method were 54%/59%/58%. In (Sourina and Liu, 2013), an emotion assessment of EEG data from which Fractal Dimension and Higher Order Crossings features have been extracted conducts to a classification performance of 53.7% for the recognition of up to 8 emotions and 87% for the recognition of 2 emotions, using 4 electrodes. For affective elicitation, in these experiments the users have been presented emotional-stimulating videos, as well as images and sounds from the IAPS and IADS databases (Sourina and Liu, 2013). Chanel et al (Chanel et al, 2011) proposed a method for adapting game levels difficulty, according to the user's emotional states – boredom, engagement and anxiety. The best accuracy has been obtained by employing ANOVA as feature

selection method and LDA as classifier. Marin-Morales et al (Marin-Morales et al, 2018) designed four virtual immersive environments with varying levels of colour, illumination and geometry with the purpose of eliciting the 4 possible combinations of arousal-valence from the Circumplex Model of Affects. EEG and ECG relevant features have been extracted using the PCA algorithm and the SVM classifier predicted with an accuracy of 71% and 75% along the valence, respectively, arousal dimensions. Although many machine learning techniques have been employed for classification, SVM remains the most popular method, together with PSD features extracted from the frequency bands of the EEG signal (Nafjan et al, 2017). For feature dimensionality reduction, SFS appears to be the most adopted approach (Bontchev, 2016).

Recently, deep learning approaches emerged in the field of emotion recognition. Zheng and Lu (Zheng and Lu, 2015) employed Deep Belief Networks (DBF) for recognizing three emotional levels – positive, neutral and negative from differential entropy features extracted from EEG signals. The DBN model's accuracy (86.08%) exceeds that of shallow models (SVM – 83.99%, LR- 82.7% and KNN- 72.6%). In the study presented in (Jirayucharoensak et al, 2014), a deep network with a stacked autoencoder is used to discriminate 3 levels of valence and arousal from 32-channel PCA dimensionally reduced EEG data. The classification rate is 53.42% and 52.05%. With SVM, the classification is 41.12% and 39.02%.

Alhagry et al (Alhagry et al, 2017) used a Long-Short Term Memory network to classify raw EEG signals from the DEAP database into low/high arousal, valence and linking with accuracies of 85.65%, 85.45% and 87.99%. In human-centric emotion recognition and affective assessment experiments, classification accuracy depends on the context of the experiment, pursued objectives, methodology, biophysical data recording procedure, number of users, structure and cleanness of training dataset, feature extraction methods, cross-validation approach and classifier statistical power & parameters tuning.

Classification accuracy depends on the context of the experiment, pursued objectives, methodology, biophysical data recording procedure, number of users, structure and cleanness of training dataset, feature extraction methods, cross-validation approach and classifier statistical power & parameters tuning. *Our approach detaches from these previous research methods, as it adds elements of novelty and originality that consist in an automatic prediction of the next game difficulty level using a trained deep neural network.*

5 The Acrophobia VRET Game. A Deep Neural Networks Approach

In the proposed VR system, a game level is characterized by a degree of exposure to a certain height. This level is selected in real-time to ensure that the player faces a challenging scenario, without forcing him into an extreme situation. We recorded in real-time the EEG and biophysical data of the players and used *two DNNs: one for fear level classification (DNN1) and one for determining the next level of the game, according to the desired level of fear (DNN2)*. This sequence of events is included in a game epoch. We call an epoch the execution of the game at a certain level.

Figure 1 presents the system's architecture and workflow. The user interacts with the Acrophobia VRET and the Virtual Game. Using the Data Acquisition Module, his EEG, HR and GSR signals are collected, pre-processed and transferred to the Database Management System (DBMS). At each game epoch, the user provides his self-assessed fear level of the current game level, called Subjective Unit of Distress (SUD). The SUD is used to determine the prediction accuracy of DNN1. The EEG, HR and GSR data are fed to DNN1 to determine the current fear level. Based on the EEG, biophysical data and desired fear level, DNN2 predicts the game level to be played next.

For fear level prediction, we used two different fear level scales. *For the **2-choices scale**, 0 represents relaxation and 1 stands for fear. For the **4-choices scale**, 0 is mapped to complete relaxation, 1 to low fear, 2 to moderate fear and 3 to a high level of anxiety.*

In order to determine the next game level to be played, we used the following approach: we considered n ordered game levels, each level corresponding to a degree of height exposure: l_0, l_1, … l_{n-1}. The user starts playing the first level of the game (l_0). The current game level is l_{cr}. During the game, the EEG

and biophysical data are recorded. When a level is completed, the biophysical averaged values are computed and using the trained DNN1, the current fear level (fl_{cr}) is determined. To achieve a gradual and appropriate exposure to height, the next desired fear level (fl_d) is calculated.

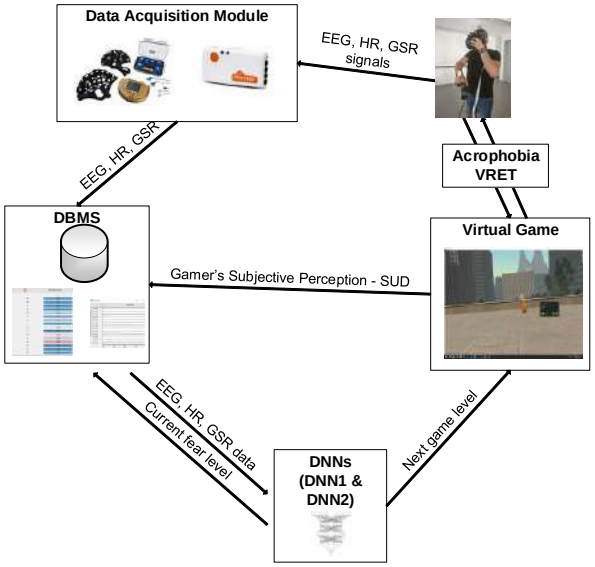

Figure 1. System architecture and workflow

There are two situations: one that considers the 2-choices scale, another that considers the 4-choices scale.

For the 2-choices scale we applied the formulas:

1. if $fl_{cr} = = 0$ then $fl_d = 1$
2. if $fl_{cr} = = 1$ then $fl_d = 0$

Thus, if the user experiences no fear at all (fl_{cr} is 0), we want to move him to a more challenging level, so the desired fear level will output 1, which means a certain level of fear. On the other hand, if the player experiences fear (fl_{cr} is 1), then he may find himself in a too difficult situation during height exposure and then we reduce the desired fear level to 0.

For the 4-choices scale we applied the formulas:

3. if $fl_{cr} = = 0$ or $fl_{cr} = = 1$ then $fl_d = fl_{cr}+1$
4. if $fl_{cr} = = 2$ then $fl_d = fl_{cr}$
5. if $fl_{cr} = = 3$ then $fl_d = fl_{cr} - 1$

If the user records complete relaxation (fl_{cr} is 0) or a low level of fear (fl_{cr} is 1), then we want to move him to a more difficult level, corresponding to a higher intensity of height exposure. Thus, we calculate the desired fear level to be one level higher than the current one. If the player currently experiences complete relaxation, we want him in the next game level to experience a low level of fear (fl_d will be 1). If in the current game level he feels low fear, we want him to go through a medium anxiety intensity in the next game level (fl_d will be 2). Moreover, if fl_{cr} is 2, corresponding to a medium level of fear, we maintain the desired fear level to this score, as it means that the player is neither too relaxed nor too anxious and the game level is challenging enough to ensure a motivating and exciting gameplay experience with appropriate height exposure. On the other hand, in the situation when the current fear level has a value of 3, pointing to extreme fear, then the desired fear level will be reduced to 2 – medium fear level – so that the prediction algorithm will take the player to a lower game level where height exposure will be adequate to meet his emotional characteristics.

The desired fear level and biophysical data are inputs for the second deep neural network (DNN2) and a game level (l_{pr}) is predicted to be played next by the user. Consequently, the user plays the predicted

level of the game and his EEG and physiological data are recorded. DNN1 determines again a new general fear level and DNN2 predicts the next game level to be played. The process goes on until a total predefined number of epochs is reached. Figure 2 presents the game workflow for the 4-choices scale.

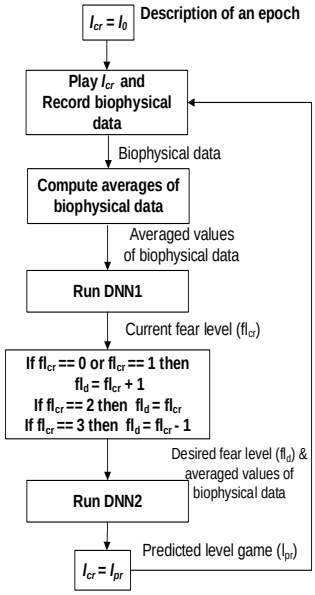

Figure 2. Description of game workflow for the 4-choices scale

6 Method and experimental results

We performed an experiment in which 4 volunteer subjects - 3 women and 1 man, aged 21-49, who have previously been informed about the purpose of the experiment and signed a consent form, played the acrophobia game while their EEG and biophysical data have been recorded. *The in-game height level exposure was predicted in real-time by DNN2, based to the user's fear score (estimated by DNN1), EEG, HR and GSR data*.

6.1 The DNN models and their cross-validation accuracies

In order to **train the deep neural networks DNN1 and DNN2**, we performed some preliminary experiments in which we gradually exposed the subjects to different heights, in both the real and virtual world. In the real-world, some baseline measurements have been performed during complete relaxation, as well as at the first, fourth and sixth floors of a building, at about 4m, 2m and a few centimetres away from the balcony's railing. Each user has been in-vivo exposed to these height levels twice, before and after the virtual exposure. EEG and biophysical data have been recorded (GSR and HR), as well as the user's perceived level of fear, the Subjective Unit of Distress (SUD). The SUD was recorded on the *11-choices scale*, where the subject had to indicate his self-assessed fear level on a scale from 0 to 10, where 0 represents complete relaxation and 10 stands for extreme fear, anxiety and panic attack. In the case of the in-vivo experiments, the SUD was reported verbally to the researcher assisting the experiment, while for the virtual exposure, the user indicated the SUD by pointing a virtual laser with the controller on a panel that appeared in from of him in the virtual environment.

For recording EEG data, we used the Acticap Xpress Bundle (Acticap Xpress Bundle) device with 16 dry electrodes, where the ground and reference electrodes have been attached to the ears. The dry electrodes have better conductance and are more comfortable for the patient than the wet ones. The following positions have been used, according to the 10/20 system: FP1, FP2, FC5, FC1, FC2, FC6, T7, C3, C4, T8, P3, P1, P2, P4, O1, O2. We recorded the alpha, beta and theta log-normalized powers for all channels, as well as the ratio of the theta to the beta powers (slow waves/fast waves). Electrodermal activity

and heart rate values have been recorded using the GSR unit of the Shimmers Multi-Sensory device (Shimmer Sensing) that was attached to the subject's left hand.

In the virtual environment, the patients had to collect coins of different colours (bronze, silver and gold) at the ground level, first, fourth, sixth and eighth floors of a building. For providing an immersive and interactive experience, the game has been integrated with the HTC Vive head-mounted display. The player perceived the environment via the virtual glasses, while the interaction was guaranteed by using the controllers – clicking on the floor at various positions by pressing the central button of the controller ensured teleportation, while the coins could be grabbed by pressing and releasing the hair trigger button. Each user played the game twice. *In both virtual and real-world conditions, we totalized a number of 63 trials per patient. These trials resulted in a dataset of 25 000 entries (or data vectors) on average for each patient, recorded at intervals of 65 ms*. The datasets were saved in .csv files and used for training DNN1 and DNN2. Thus, although we have a small number of users, we still benefit from a large training dataset to be fed to the networks in order to create reliable prediction models.

Our first goal was to create an accurate and reliable deep neural network model (DNN1) for estimating fear level, based on the recorded EEG and biophysical data. For training DNN1, we used the data recorded during the 63 trials in both the real-world and virtual environment. *The input features were the EEG log-normalized powers of all the channels, GSR and HR values, whereas the output feature was the fear level (or SUD), on the 11-choices (values from 0 to 10), 2-choices (values of 0 or 1) and 4-choices (values from 0 to 3) scales*. The 2-choices and 4-choices scales have been obtained by grouping together the values from the 11-choices scale. For the 11-choices-scale, 0 corresponded to complete relaxation, values from 1-3 to small levels of fear, from 4-7 to medium levels of fear and 8-10 to high anxiety. For the 4-choices-scale, the previous scale values have been grouped together, so that finally 0 corresponded to complete relaxation, 1 to small, 2 to medium, 3 to high level of fear. On the 2-choices-scale, the 4-choices-scale values have been also grouped together, in 0 (for the first two levels, meaning relaxation) and 1 (for the third and fourth levels, meaning fear) values. This grouping has been done in order to improve categorization and classification in the neural networks.

Using the TensorFlow (Tensor Flow Python Framework) deep learning framework backend, we created four Keras (Keras library) sequential models for binary and multi-class classification: Model_1, Model_2, Model_3 and Model_4.

Model_1 has 3 hidden layers, with 150 neurons on each layer. Model_2 has 3 hidden layers, with 300 neurons on each layer. Model_3 has 6 hidden layers, with 150 neurons on each layer. Model_4 has 6 hidden layers, with 300 neurons on each layer. For all the models, the hidden layers used the Rectified Linear Unit (RELU) activation function. For the 2-choices scale, we used the Sigmoid activation function in the output layer and the binary crossentropy loss function. However, for the 4-choices and 11-choices scales, we used the Softmax activation function in the output layer, logarithmical categorical crossentropy loss function and one-hot-encoding that creates 4 or 11 output values (correspondingly 4 or 11 neurons), one for each class. The largest output value will be taken as the class predicted by the model. The models also employ the efficient Adam gradient descent optimization algorithm. Prior to training, the data has been standardized, to reduce it to zero mean and unit variance. The Keras Classifier received as arguments a number of 1000 epochs for training and a batch size of 20. The neural network model has been evaluated on the training data using the KFold method from the scikit-learn library (Scikit Learn Python Library).

We ran the evaluation using a 10-fold cross-validation procedure 10 times and saved the weights of the network in .hdf5 files, together with the corresponding accuracies. Finally, the model version with the highest accuracy has been selected and further used in the experiment. This procedure has been repeated for each user, so that *we obtained four personalized fear estimation DNN1 models for each user, for the 2-choices, 4-choices and 11-choices fear scales*.

Our second goal was to define a deep neural network model (DNN2) to predict the next game level. *For training DNN2, the neural network received as inputs the EEG, GSR, HR and SUD values, while the output represented an encoding of the height where these physiological values have been recorded – 0 for ground floor, 1 for the first floor, 2 for the fourth floor, 3 for the sixth floor and 4 for the eighth floor*.

Similarly, using TensorFlow and Keras, we created four multilayer perceptron sequential models – Model_1, Model_2, Model_3, Model_4, RELU activation function for the hidden layers, Softmax activation function for the output layer, logarithmical categorical crossentropy loss function and one-hot-encoding that creates 5 output values (correspondingly 5 neurons), one for each estimated class. Model_1 has 3 hidden layers, with 150 neurons on each layer. Model_2 has 3 hidden layers, with 300 neurons on each layer. Model_3 has 6 hidden layers, with 150 neurons on each layer. Model_4 has 6 hidden layers, with 300 neurons on each layer. The Keras Classifier received the same arguments and the 10 times evaluation with the 10-fold cross-validation procedure was similar to DNN1. The procedure has been repeated for each user, so that *we obtained a personalized height exposure (game level) DNN2 model for each user, for both the 2-choices, 4-choices and 11-choices scales.* Similarly to DNN1, the model version with the highest accuracy has been selected and further used in the experiment. The maximum cross-validation accuracies obtained for the models are presented in Table 1.

DNN1			
Model	Maximum cross-validation accuracy (%)		
	2-choices scale	*4-choices scale*	*11-choices scale*
Model_1	95.03	87.94	85.09
Model_2	95.51	90.49	79.48
Model_3	94.43	86.32	74.27
Model_4	94.57	88.28	80.45
DNN2			
Model	Maximum cross-validation accuracy (%)		
	2-choices scale	*4-choices scale*	*11-choices scale*
Model_1	98.40	98.67	98.75
Model_2	98.72	98.50	98.65
Model_3	97.45	97.82	98.50
Model_4	97.37	97.77	98.17

Table 1. Maximum cross-validation accuracies for DNN1 and DNN2

Besides deep neural networks with the 4 architectural models from above, we also trained our data using the Linear Discriminant Analysis classifier. We obtained a cross-validation accuracy of 87% for the 2-choices model, 71% for the 4-choices model and 64% for the 11-choices model.

6.2 The Game Development

The VR-based game has been developed using the Unity game engine (Unity Game Engine) and written entirely in the C# programming language. It has integrated connectivity with the OpenVibe (OpenVibe) application for collecting EEG signals and with Shimmers Capture (Shimmers Capture C# API) for recording GSR and HR data. The Shimmers Capture application has been modified according to our game integration needs, thanks to the availability of its C# API. The recordings were synchronized in real-time using Lab Stream Layer (LSL) (Lab Stream Layer). The game starts with the user placed on the ground floor (Figure 3). After he collects three coins (bronze, silver and gold), the application remains in standby for a few seconds, the EEG and biophysical data are averaged and the fear level & next game level prediction processes take place in the background. Thus, the corresponding Python scripts for testing are selected, according to the current user and condition (2-choices scale or 4-choices scale). The 11-choices scale has not been integrated yet. *First, the script for fear level estimation (DNN1) is called and predicts the current fear level. The desired fear level is calculated based on the formulas described in Chapter 5. Secondly, the script for next game level (or next height exposure level) estimation is called (DNN2), predicting the level where the user should be taken in the next gameplay epoch*.

We introduced some gamification elements, i.e. the challenge of collecting coins, as it adds interactivity and purpose to gameplay. The coins are placed at gradual distances from the building's balcony railing, so that for collecting the golden one it requires the user to bend over the railing and forcefully catch a glimpse of the view (Figure 4).

Figure 3. Screenshot of the virtual environment from the ground floor

Figure 4. Screenshot of the virtual environment from the fourth floor of the building

For each epoch, the averaged EEG, GSR and HR values are stored in log files, together with the predicted fear and game levels. Prior to entering the DNNs, the data is denoised and pre-processed. As the recording devices introduce noise, interrupt temporarily, disconnect or malfunction, we applied a method called "last good value". For instance, if the HR value at a moment of time is invalid (a negative or a very big number, which is a clear sign of failure), we replace it with the last good value recorded at a previous timestamp (let's say 86 bpm). If the device malfunction from the beginning, we initialize the last good value with 4.5 microVolts^2 for the EEG log-normalized power, 1 microSiemens for GSR and 75 bpm for HR. We applied this method because our application runs in real-time and we are not able to manually interfere for inspecting, interpolating or removing the noisy data. Even though we used advanced and expensive sensory devices, the drawback of being unable to fully rely on the recording tools still persists.

Moreover, for each epoch, we saved in separate log files the EEG alpha, beta, theta, GSR and HR values, recorded at intervals of 65 miliseconds. They are saved in both unprocessed and processed (denoised) format, being useful for further experimentation and analysis.

6.3 Experiment and results

Our 4 subjects played the game twice – once using the 2-choices and once using the 4-choices model. Each session contained a number of 10 game epochs. After the user finished one epoch and succeeded in collecting the three coins, he was required to report the perceived fear level for that particular trial (the SUD). A menu appeared on the screen and the answer was given by pointing to the value corresponding to the current self-estimated SUD. Consequently, his biophysical data was saved, together with the current SUD and DNN1 & DNN2 started to run in the background for establishing the current fear level and the next game level where the player should be automatically taken. The purpose of collecting self-estimated SUDs was to validate the accuracy of DNN1. DNN1 predicted the current fear level based on a neural network model created using the data from the previous experimentation and a measure of certifying its faultlessness was by comparing its output with the fear level perceived and acknowledged by the users directly during gameplay – the SUD. We called this parameter *validation accuracy*. The validation accuracies for each model are presented in Table 2.

Model	Validation accuracy for DNN1 (%)	
	2-choices scale	*4-choices scale*
Model_1	72.90	41.89
Model_2	68.73	24.99
Model_3	62.45	34.15
Model_4	54.12	38.32

Table 2. Validation accuracy for DNN1

The validation accuracy of the LDA classifier for the same data is 60% for the 2-choices scale and 21% for the 4-choices scale. We conclude that Model_1 provided the best training cross-validation and test validation accuracies for DNN1 – for the 2-choices scale, a cross-validation accuracy of 95.03% and validation accuracy of 72.90%. For the 4-choices scale, the values are of 87.84% and 41.89%. As we have not designed yet a method for validating whether the game levels predicted by DNN2 in the experiment are appropriately determined, we do not have a test set for DNN2 and thus we could not calculate its validation accuracy. The only modality used for assessing the efficiency of the proposed approach was by comparing the SUDs reported by the subjects during gameplay with the fear score predicted by DNN1.

The test set is small, containing 10 records for each game session played, this is probably the reason why the validation accuracy did not reach a higher value, especially for the 4-choices scale. Moreover, User1 recorded a low validation accuracy for both the 2-choices and the 4-choices scale, whereas the other users obtained a validation accuracy of over 75%. Due to the poor test results of User1, the average validation accuracy for all the users dropped to the values of approximately 73%, respectively 42%, as presented in Table 2. Without taking into account the data from User1, the validation accuracy of DNN1 is 85% for the 2-choices scale and 60% for the 4-choices scale.

The game levels varied throughout gameplay according to the fear scores, with good results for 3 out of 4 users who recorded a level of fear of 0 (complete relaxation) in the final gameplay epoch. The fourth subject, who suffered from a more severe form of acrophobia, recorded a fear level of 2 in the final game epoch.

A Dynamic Difficulty Adjustment (DDA) of game levels based on the affective state information was proposed by Liu et al (Liu et al, 2009), with prediction accuracy of 78%. Our method conducted to a comparable accuracy. However, their adjustment was based on some simple "if" clauses, not on an advanced prediction method, as ours. Chanel et al (Chanel et al, 2011) tried to adapt the game difficulty levels to the players' emotional states (boredom, engagement and anxiety). Without feature selection, the best classifiers obtained an accuracy of 55% for peripheral signals and 48% for EEG (LDA, followed by SVM). After the fusion of the two signal categories, their accuracy increased to 63%. We conclude that our classifiers performed equally good, with accuracies of 73% and 42%. The results are promising, but in order to demonstrate the strengths of our method, more experiments need to be done and with a larger number of subjects.

7 Conclusions and future work

This paper presented a real-time deep learning - based approach for treating acrophobia in the virtual environment. Two complex neural networks that have been trained with the subjects' data from an experimental procedure where they have been in-vivo and virtually exposed to different heights. The same users participated in an experiment where they were required to play the game, have their EEG and biophysical signals recorded, report the perceived fear level, but advance to the next game automatically, based on the output provided by the two neural networks. The validation accuracy, defined as the measure of similarity between the fear level estimated by the first deep neural network and the fear level reported subjectively by the user was 73% and 42%. The game levels varied throughout gameplay according to the relaxation / anxiety scores, with good results for 3 out of 4 users who recorded 0 level of

fear in the final gameplay epoch. The main challenges are represented by the instability of the sensory recording devices that sometimes fail to connect, introduce noise or errors in the data. It is very important to have clean data, for both training and testing the neural network models, as they can influence the prediction accuracy. In this phase, offline and online pre-processing and denoising is an essential, indispensable step.

Relying on the promising obtained results, we will continue to extend the research for other types of phobias. Moreover, we will try to use other machine learning techniques, in order to determine the best solutions and make a comparison between a totally automatic approach and a human-centred approach. In addition, we will perform more experiments with a larger number of users and do real-world tests in order to validate the efficiency of the VR treatment and see whether their acrophobic condition has indeed improved.

Acknowledgment

This work has been funded by UEFISCDI proiect 1/2018, UPB CRC Research Grant 2017 and UEFISCDI project PN-III-P1-1.1-TE-2016-2182.

References

2BIOPAC. URL: https://www.biopac.com/application/virtual-reality/ (visited on 20/11/2018).

Acticap Xpress Bundle. URL: https://www.brainproducts.com/productdetails.php?id=66 (visited on 20/11/2018).

Alhagry, S., Fahmy, A.A., El-Khoribi, E.A. 2017. "Emotion Recognition based on EEG using LSTM Recurrent Neural Network". *Int. J. Adv. Comput. Sci. Appl*. 8, 355–358

AlZoubi, O., D'Mello, S.K., Calvo, R.A. (2012). "Detecting naturalistic expressions of non-basic affect using physiological signals". *Affective Computing, IEEE Transactions on*, 3(3):298–310.

Arikan, K., Boutros, N.N., Bozhuyuk, E., Poyraz, B.C., Savrun, B.M., Bayar, R., et al (2006). "EEG correlates of startle reflex with reactivity to eye opening in psychiatric disorders: preliminary results". *Clin EEG Neurosci*., 37:230-4.

Bălan, O., Moise, G., Moldoveanu, A., Leordeanu, M., Moldoveanu, F. (2019). "Challenges for ML-based Emotion Recognition Systems in Medicine. A Human-Centered Approach ". CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

Bălan, O., Moise, G., Moldoveanu, A., Moldoveanu, F., and Leordeanu, M. (2018). "Does automatic game difficulty level adjustment improve acrophobia therapy? Differences from baseline.". *In Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (VRST '18), Stephen N. Spencer (Ed.). ACM, New York, NY, USA, Article 78, 2 pages. DOI: https://doi.org/10.1145/3281505.3281583

Bontchev, B. (2016). "Adaptation in affective video games: A literature review". *Cybern. Inf. Technol*. 16, 3–34

Bos, D.O. (2006). "EEG-based emotion recognition. The Influence of Visual and Auditory Stimuli"

C2Phobia. URL: https://www.c2.care/en/c2phobia-treating-phobias-in-virtual-reality/ (visited on 20/11/2018).

Canli, T., Desmond, J.E., Zhao, Z., Glover, G., Gabrieli, J.D.E. (1998). "Hemispheric asymmetry for emotional stimuli detected with fMRI". *NeuroReport*, 9(14), 3233–3239.

Chanel, G., Rebetez. C., Bétrancourt, M., and Pun, T. (2011). "Emotion Assessment from Physiological Signals for Adaptation of Game Difficulty". *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 41(6), 1052-1063.

Davidson, R.J. (1993). "Cerebral asymmetry and emotion: Conceptual and methodological conundrums". *Cognit. Emot*., 7: 115–138.

Demaree, H. A.; Everhart, D. E.; Youngstrom, E. A.; Harrison, D. W. 2005. "Brain lateralization of emotional processing: Historical roots and a future incorporating dominance.". *Behavioral and Cognitive Neuroscience Reviews*. 4, 3– 20

DiMeglio, C. (2015). "Fear Feedback Loop: Creative and Dynamic Fear Experiences Driven by User Emotion". Master Thesis, Rochester Institute of Technology

Eaton, W.W., Bienvenu,O.J., Miloyan, B., (2018). "Specific phobias". *Lancet Psychiatry*. 5(8), 678-686. doi: 10.1016/S2215-0366(18)30169-X

Fleureau, J., Philippe, G., Huynh-Thu, Q. (2012). "Physiological-based affect event detector for entertainment video applications". *Affective Computing, IEEE Transactions on*, 3(3), 379–385.

Freeman, D., Haselton, P., Freeman, J., Spanlang, B., Kishore, S., Albery, E., Denne, M., Brown, P., Slater, M., Nickless, A., (2018). "Automated psychological therapy using immersive virtual reality for treatment of fear of heights: a single-blind, parallel-group, randomised controlled trial". *Lancet Psychiatry*, 5: 625–32, Published Online, July 11, 2018, http://dx.doi.org/10.1016/

Garcia-Palacios, H. G., Hoffman, S., Kwong See, A., Botella, C. (2001). "Redefining Therapeutic Success with Virtual Reality Exposure Therapy". *CyberPsychology & Behavior*, 4(3), 341–348, http://online.liebertpub.com/doi/abs/10.1089/109493101300210231

Healey, J. (2009). "Affect detection in the real world: Recording and processing physiological signals". *Affective Computing and Intelligent Interaction and Workshops*, 2009. ACII 2009. 3rd International Conference on, pages 1–6. IEEE

HTC Vive. URL: https://www.vive.com/eu/ (visited on 20/11/2018).

Institute of Health Metrics and Evaluation (IHME), Global Burden of Disease (GBD) Deaths, DALYs and prevalence of mental health and substance use disorders, by age and sex, Geographical coverage: Global by country and region, Time span: 1990 – 2016, URL: http://ghdx.healthdata.org/gbd-results-tool (visited on 3/12/2019).

International Classification of Diseases ICD (2018), ICD-11 for Mortality and Morbidity Statistics, URL: https://icd.who.int/browse11/l-m/en (visited on 3/12/2019).

Jirayucharoensak, S., Pan-Ngum, S., Israsena, P. (2014). "EEG-based emotion recognition using deep learning network with principal component-based covariate shift adaptation". *The Scientific World Journal*.

Jones, N.A., Fox, N.A., (1992). "Electroencephalogram asymmetry during emotionally evocative films and its relation to positive and negative affectivity". *Brain and Cognition*, 20(2), 280–299.

Keras Library. URL: https://keras.io/ (visited on 20/11/2018).

Koelstra, S., Muehl, C., Soleymani, M, Lee, J.-S, Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I. (2012). "DEAP: A Database for Emotion Analysis using Physiological Signals". *IEEE Transactions on Affective Computing*, 3, 18-31

Koelstra, S., Yazdani, A., Soleymani, M., Muhl, C., Lee, J.-S., Nijholt, A., Pun, T., Ebrahimi, T., Patras, I. (2010). "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos". *Brain Informatics, ser. Lecture Notes in Computer Science*, 6334(9), 89–100

Kometer, H., Luedtke, S., Stanuch, K., Walczuk, S., Wettstein, J. (2010). "The Effects Virtual Reality Has on Physiological Responses as Compared to Two-Dimensional Video", University of Wisconsin School of Medicine and Public Health, Department of Physiology, (2010)

Lab Stream Layer. URL: https://github.com/sccn/labstreaminglayer (visited on 20/11/2018).

Levski, Y. 15 Greatest Examples of Virtual Reality Therapy, URL: https://appreal-vr.com/blog/virtual-reality-therapy-potential/, (visited on 3/12/2019).

Liu, C., Agrawal, P., Sarkar, N. and Chen, S. (2009). "Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback". *International Journal of Human–Computer Interaction*, 25(6), 506-529.

Marín-Morales, J., Higuera-Trujillo, J.L., Greco, A., Guixeres, J., Llinares, C., Scilingo, E.P. (2018). "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors". *Sci. Rep*. 8:13657. doi: 10.1038/s41598-018-32063-4

Nafjan, A. et al. (2017). Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review. *Appl.Sci*. 7, 1239. DOI:10.3390/app7121239

North, M.M., North, S.M., and Joseph R. Coble, J. R. (1997). "Virtual Reality Therapy: An Effective Treatment for Psychological Disorders". *Virtual Reality in Neuro-Psycho-Physiology*, Giuseppe Riva (Ed.), Ios Press: Amsterdam, Netherlands

Olesen, J., (2015). *Phobia Statistics and Surprising Facts About Our Biggest Fears*. URL: http://www.fearof.net/phobia-statistics-and-surprising-facts-about-our-biggest-fears/ (visited on 20/11/2018).

OpenVibe. URL: http://openvibe.inria.fr/ (visited on 20/11/2018).

Opris, D., Pintea, S., Garcia-Palacios, A., Botella, C., Szamoskozi, S., David, D. (2012). "Virtual Reality exposure therapy in anxiety disorders: a quantitative meta-analysis". *Depress Anxiety*, 29, 85-93.

PSIOUS. URL: https://www.psious.com/ (visited on 20/11/2018).

Ritchie, H. and Roser, M. "Mental Health". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/mental-health (visited on 3/12/2019).

Ritchie's Plank Experience. URL: http://store.steampowered.com/app/517160/Richies_Plank_Experience/ (visited on 20/11/2018).

Robertson, A. (2016). *The Climb turns virtual reality acrophobia into an extreme sport*. URL: https://www.theverge.com/2016/4/28/11526150/crytek-the-climb-vr-oculus-rift-review (visited on

Russell, J.A. (1979). "Affective space is bipolar". *Journal of Personality and Social Psychology*, 37(3), 345–356

Schafer, P., Koller, M., Diemer, J., Meixner, G. (2015). "Development and evaluation of a virtual reality-system with integrated tracking of extremities under the aspect of Acrophobia". *SAI Intelligent Systems Conference (IntelliSys)*, 408–417. IEEE, London

Scikit Learn Python Library. URL: http://scikit-learn.org (visited on 20/11/2018).

Shimmer Sensing. URL: http://www.shimmersensing.com/ (visited on 20/11/2018).

Shimmers Capture C# API. URL: http://www.shimmersensing.com/products/shimmercapture (visited on 20/11/2018).

Soleymani, M., Kierkels, J., Chanel, G., Pun, T. A. (2009). "Bayesian framework for video affective representation". Proc. Int. Conf. Affective Computing and Intelligent interaction.

Sourina, O., Liu, Y., Nguyen, M.K. (2012). "Real-time EEG-based emotion recognition for music therapy". J. *Multimodal User Interf*, 5(1–2), 27–35

Tensor Flow Python Framework. URL: https://www.tensorflow.org/ (visited on 20/11/2018).

Trainor, L.J., Schmidt, L.A. (2003). "Processing Emotions Induced by Music". Cognitive Neuro-science of Music (Oxford), 317p.

Unity Game Engine. URL: https://unity3d.com/ (visited on 20/11/2018).

Virtual Reality Medical Center. URL: http://www.vrphobia.com/aboutus.htm. (visited on 20/11/2018).

Wardenaar, K.J. et al (2017). *The cross-national epidemiology of specific phobia in the World Mental Health Surveys*. URL: https://repositori.upf.edu/bitstream/handle/10230/34033/Wardenaar-psm-thec.pdf?sequence=1&isAllowed=y (visited on 20/11/2018).

Westerink, J., Ouwerkerk, M., de Vries, G.J., de Waele, S., van den Eerenbeemd, J., van Boven, M. (2009). "Emotion measurement platform for daily life situations". *In Affective Computing and Intelligent Interaction and Workshops*, 2009.

WHO (2017). Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization; 2017. (visited on 3/12/2019).

World Health Organization (WHO) Global Health Observatory (GHO), Prevalence of depression, Geographical coverage: Global by country, Time span: 2015, URL: http://apps.who.int/gho/data/node.home. (visited on 3/12/2019).

Zheng, W.-L. and Lu, B.-L. (2015). "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks". IEEE Transactions on Autonomous Mental Development, 7(3), 162–175

sensors

MDPI

Article

An Investigation of Various Machine and Deep Learning Techniques Applied in Automatic Fear Level Detection and Acrophobia Virtual Therapy

Oana Bălan [1,*], Gabriela Moise [2], Alin Moldoveanu [1], Marius Leordeanu [1] and Florica Moldoveanu [1]

1 Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, Bucharest 060042, Romania; alin.moldoveanu@cs.pub.ro (A.M.); marius.leordeanu@cs.pub.ro (M.L.); florica.moldoveanu@cs.pub.ro (F.M.)

2 Department of Computer Science, Information Technology, Mathematics and Physics, Petroleum-Gas University of Ploiesti, Ploiesti 100680, Romania; gmoise@upg-ploiesti.ro

* Correspondence: oana.balan@cs.pub.ro; Tel.: +4072-2276-571

Received: 29 October 2019; Accepted: 13 January 2020; Published: 15 January 2020

Abstract: In this paper, we investigate various machine learning classifiers used in our Virtual Reality (VR) system for treating acrophobia. The system automatically estimates fear level based on multimodal sensory data and a self-reported emotion assessment. There are two modalities of expressing fear ratings: the 2-choice scale, where 0 represents relaxation and 1 stands for fear; and the 4-choice scale, with the following correspondence: 0—relaxation, 1—low fear, 2—medium fear and 3—high fear. A set of features was extracted from the sensory signals using various metrics that quantify brain (electroencephalogram—EEG) and physiological linear and non-linear dynamics (Heart Rate—HR and Galvanic Skin Response—GSR). The novelty consists in the automatic adaptation of exposure scenario according to the subject's affective state. We acquired data from acrophobic subjects who had undergone an in vivo pre-therapy exposure session, followed by a Virtual Reality therapy and an in vivo evaluation procedure. Various machine and deep learning classifiers were implemented and tested, with and without feature selection, in both a user-dependent and user-independent fashion. The results showed a very high cross-validation accuracy on the training set and good test accuracies, ranging from 42.5% to 89.5%. The most important features of fear level classification were GSR, HR and the values of the EEG in the beta frequency range. For determining the next exposure scenario, a dominant role was played by the target fear level, a parameter computed by taking into account the patient's estimated fear level.

Keywords: fear classification; emotional assessment; feature selection; affective computing

1. Introduction

According to statistics, 13% of the world's population is affected by phobias, a type of anxiety disorder manifested by an extreme and irrational fear towards an object or a situation. 275 million people suffer from anxiety disorders throughout the world and anxiety disorders are ranked as the 6th-most common contributors to global disability [1]. Phobias are classified into social phobias (fear of relating to others or speaking in public) and specific phobias (generated by particular objects or situations). Social phobias affect people of all ages, though they usually start to manifest in adolescence. 17% of people with social phobias develop depression. The majority of them turn to medication, and even substance abuse and illegal drugs (nearly 17%) or alcohol (nearly 19%), and only 23% seek specialized help [2]. With regard to specific phobias, a significant percent (15–20%) of the world's population faces one specific phobia during their lifetime [3]. The most common specific phobias

and their prevalence are: acrophobia (fear of height)—7.5%; arachnophobia (fear of spiders)—3.5%; aerophobia (fear of flying)—2.6%; astraphobia (fear of lightning and thunder)—2.1%; and dentophobia (fear of dentist)—2.1%. [4]. Specific phobias begin during childhood and can persist throughout one's life, affecting more women than men. Most of these patients do not seek treatment for phobias and, of those who do, only 20% recover completely [2].

The treatment for phobias is either medical or psychological. 80% of people suffering from phobias turn to medicines and Cognitive Behavior Therapy (CBT), a form of psychotherapy that encourages patients to modify destructive patterns of cognition and behavior and to replace them with positive thoughts [5]. Immersion therapy consists of gradual exposure to anxiety-producing stimuli, in the presence of the therapist who controls the intensity of immersion [6]. Thus, the patients are urged to understand their fears and find a way to adjust their attitude towards the anxiety-provoking object/situation in a conscious and apperceptive fashion. The medical or psychological treatment should be continued for as long as required since statistics reveal that phobia tends to relapse in approximately 50% of cases [7]. With the technological advancement, Virtual Reality has significantly emerged in recent decades, allowing the design of immersive virtual worlds that provide stimuli in a safe and controlled manner [8].

In 1997, Picard published a seminal book entitled Affective Computing, in which are presented the theories and principles of a new interdisciplinary field encompassing computer science, neuroscience, psychology, and engineering [9]. Affective Computing (AC) is defined as "computing that relates to, arises from, or influences emotions".

According to Picard, computers need to understand human emotions and even have and express emotions for the purpose of communicating with humans. AC enables an integration of human emotions into technology. The field comprises: the study of affect recognition and generation methods, expressing affection techniques, affect aware systems development, research on the modality in which affect influences human-technology interactions. AC helps people understand psychological phenomena, human behaviors, and to build better software applications [10]. AC has many applications in education, game development, health, robots, cyber-psychology, VR, marketing, entertainment, and so on.

The integration of affective information in game development opens the path to new methods of maintaining players' engagement [11], by dynamically adjusting game levels difficulty to tailor the users' individual emotional characteristics [12]. In healthcare applications, AC involves automatic emotion detection and provides decisions accordingly. Relational agents have been developed in order to help patients in hospitals or to assist childbirth, offering information and emotional support [13]. Conversational agents and robots interact with children suffering from ASD, helping them to develop from the socio-emotional point of view [14].

In this paper, we propose a VR game for treating acrophobia, based on the idea of real-time automatic adaptation of in-game height exposure according to the subject's level of fear. With physiological signals as input (EEG, GSR and HR), our system determines the subject's current fear level and predicts the next exposure scenario.

The current fear level and the next exposure scenario were obtained using various machine learning (ML) and deep learning (DL) classifiers: Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Linear Discriminant Analysis (LDA), Random Forest (RF), and 4 deep neural network architectural models. The data used for training the classifiers was recorded in a preliminary experiment in which 8 acrophobic subjects were in vivo and virtually exposed to various heights. For computing the accuracy of the classifiers, both a user-dependent and a user-independent modality were used. Therefore, each classifier was trained using the data of the other subjects in the case of the user-independent modality. We calculated cross-validation and test accuracies applying the trained model on the data of the tested user. Moreover, this is the research idea towards which we are inclined, given the fact that training a classifier for every subject is an unfeasible and highly time-consuming activity. On the other hand, in the case of the user-dependent modality, for each subject, each classifier was trained using their

own data, obtained from the preliminary experiment. The trained model was then applied on the test records of the same participant. Feature selection was also computed for each classifier in order to improve generalization across subjects.

To validate our method, we performed an experiment with the same 4 acrophobic users, in which they played the proposed acrophobia game twice. The first classifier predicted the current fear level, while the second one estimated the next exposure scenario (or game level to be played next). The results showed a very high cross-validation accuracy on the training set (obtained by the kNN and RF classifiers) and good test accuracies, ranging from 42.5% (for the 4-choice scale) to 89.5% (for the 2-choice scale) (both for SVM, for the player-dependent modality). Also, we determined that the most relevant features for fear level classification were GSR, HR and the values of the EEG in the beta frequency range. For the next exposure scenario prediction, an important role was played by the target fear level.

The paper is organized as follows: Section 2 presents the state of the art regarding the current VR-based therapies, Section 3 introduces a short description of the emotional models and types of physiological signals employed in our research, Section 4 details similar experiments and the modalities in which various machine learning techniques have been used for emotion classification, Section 5 presents our acrophobia game, together with the ML and DL approach for fear level and next exposure scenario prediction, Section 6 provides an insight into the methodology for training, dataset construction and experimental procedure, while Section 7 emphasizes the results of our experiments. Finally, we discuss the research findings in Section 8 and present the conclusions and future work directions in Section 9.

2. VR-Based Phobia Therapy

Virtual Reality has been involved in phobia treatment since the 1990s. In the study presented in [15], 60 participants suffering from agoraphobia have been equally divided into two groups: a control group and an experimental group. Eight virtual environment scenes were used to expose 30 participants from the experimental group in session sequences of about 15 min. The Attitude Towards Agoraphobia Questionnaire (ATAQ) and Subjective Unit of Discomfort (SUD) were used as instruments to assess the anxiety states of the subjects. SUD means decreased over the eight sessions, from 5.66 to 2.42, indicating habituation with the agoraphobic stimuli. The results proved that agoraphobic patients can be successfully treated with VR technologies. VR technologies have manifold applications in phobia treatment, from understanding the causes of these disorders, to evaluating and treating them [16,17].

Virtual Reality Exposure Therapy (VRET) is a behavioral treatment for anxiety disorders, including phobias. The patient is immersed in a computer-generated virtual environment which presents stimuli that are dangerous in real-world situations [18]. VRET is equally as efficient as the classical evidence-based interventions (CBT and in vivo exposure), provides real-life impact, has good stability of results in time and engages the patients in the therapy as much as in vivo exposure does [8,16]. The existing VRET systems can be classified into platforms, academic research projects or experiments and mobile/desktop game applications.

2.1. Platforms

C2Phobia [19] was designed by mental health professionals, psychiatrists, psychologists and psychotherapists. Using a VR headset, the patient is gradually exposed to anxiogenic situations. The system can also be used at home, allowing the specialist to treat patients and prescribe personal exercises at a distance. C2Phobia is recognized as a medical device, a complete therapeutic software, but the developers did not disclose whether they use machine learning techniques or not.

PSIOUS [20] provides animated and live VR and Augmented Reality (AR) environments, as well as 360-degree videos for anxiety disorders, fears and phobias treatment. It offers patients monitoring capabilities, generation of automatic reports and the possibility of home training. PSIOUS contains 70 VR scenes. The developers did not disclose whether they used machine learning techniques or not.

Stim Response Virtual Reality [21] provides a wide range of virtual worlds for acrophobia, aerophobia and social phobias treatment, as well as physiological data synchronization. The VR and AR scenes change in real time, based on the subject's responses to the environment. It performs automatic data analysis.

Virtual Reality Medical Center [22] uses 3D virtual environments, biofeedback and CBT to treat phobias and anxieties (especially pre-surgical anxiety), relieve stress, and teach relaxation skills. Non-invasive physiological monitoring with visual feedback allows control for both the patient and the therapist. Virtually Better [23] is a system available only for the therapist's office which is aimed at providing therapeutic applications for treating phobias, job interview anxiety, combat-related post-traumatic stress disorders, and drug or alcohol addiction. Virtually Better has been used by the VR Treatment Program at Duke Faculty [24], where the therapist guided the participants through the environment and interacted with them through the entire event. Research studies have indicated that 6 to 12 45–50-minute-long therapy sessions were enough to achieve maximum benefit. The Bravemind system [25] was used for treating soldiers who served in Iraq and Afghanistan with anxiety disorders. It works by providing vibrotactile and olfactive sensations associated with war zones. Limbix [26] contains interactive scenes made of panoramic images and videos that can be changed by the therapist in real-time. Lastly, PHOBOS [27] was designed in consideration of CBT protocols. It provides interactive environments, gradual exposure and realistic crowd and social group dynamics simulations for treating social and specific phobias.

2.2. Academic Research Projects and Experiments

Acrophobia Therapy with Virtual Reality (AcTiVity-System—UniTyLab, Hochschule Heilbronn, Germany) [28] is played on an Android device and uses the Oculus Rift headset [29] to render 3D scenes, Microsoft Kinect [30] for motion tracking, and a heart rate sensor for measuring HR. The virtual environment contains buildings that have a walk route on the sides. A large experiment was performed in which 100 users were divided into a VR group and a control group. The participants from the VR group had to take a tour in a 10-storey office complex. All 44 subjects from the VR group who completed the six sessions of the experiment had an average reduction of 68% of their fear of heights. VR Phobias [31] presents a static virtual environment depicting the view from the balcony of a hotel. The results of an experiment in which 15 acrophobic patients were exposed to heights in vivo and virtually showed that the success rates of both procedures were similar. However, the VR exposure sessions were shorter, safer and more comfortable for the patients. The acrophobia system presented in [32] contained three virtual environments in a cityscape. The results of an experiment in which twenty-nine subjects participated and rated their fear levels in the presence of a therapist who adjusted exposure according to their affective state showed that both anxiety and avoidance levels decreased. Virtual therapy proved to be as effective as in vivo exposure to fear-provoking stimuli.

2.3. Mobile/Desktop Game Applications

Some of the most popular desktop game applications available for the Oculus Rift [29] and HTC Vive [33] headsets are The Climb [34], Ritchie's Plank Experience [35], Arachnophobia [36] and Limelight [37]. The first two try to overcome fear of heights, Arachnophobia treats fear of spiders, while Limelight puts the user in front of a crowd with changeable moods where he/she gives lectures or presentations, in order to overcome their fear of public speaking. Samsung Fearless Cityscapes [38] and Samsung Fearless Landscapes [39] are dedicated to acrophobia therapy and are rendered via Gear VR [40] glasses. Hear rate can also be monitored when they are paired with Gear S2 [41].

Most of the VR applications mentioned above do not provide any details related to the technologies and the methods used. Thus, we cannot ascertain whether ML techniques were deployed for adapting the therapy. On the other hand, we are interested in building Machine Learning-based applications tailoring therapy to the individual characteristics of each patient.

Our system detects the fear level in real time and automatically selects the next exposure scenario. By training the classifiers in a user-independent way with the data obtained from in vivo and virtual experiments, we aim to construct robust classification models that would accurately evaluate the patients' affective states and adjust the levels of exposure accordingly. Thus, we intend to provide a reliable therapeutic solution for phobia alleviation based on Virtual Reality and human-centered machine learning. Our system can be used in clinics, for home therapy and deployed on mobile devices, incorporating all the advantages of the above-mentioned systems.

3. Emotion Models and Physiological Data

3.1. Emotion Models

Emotion is defined as a feeling deriving from one's circumstances, mood or relationship with others [42]. It is a complex psycho-physiological experience generated by the conscious or unconscious perception of an object or situation [43], manifested through bodily sensations and changes in mood and behavior. The bodily sensations originate from the autonomic nervous system (increased cardiac activity, dilatation of blood vessels, involuntary changes in the breathing rate), cortex (activation of emotion-related brain areas) and are accompanied by physical expressions such as tremor, crying or running [44]. Various classifications of emotions have been proposed, from both the discrete and the dimensional perspective. One of the first categorizations identified six discrete emotions: happiness, sadness, fear, anger, disgust and surprise [45]. Consequently, the list was updated with embarrassment, excitement, contempt, shame, pride, satisfaction, amusement, guilt, relief, wonder, ecstasy, and sensory pleasure [46]. Complex emotions can be constructed from a combination of basic emotions. Plutchik introduced the wheel of emotions to illustrate how basic emotions (joy versus sadness; anger versus fear; trust versus disgust; surprise versus anticipation) can be mixed to obtain different emotions [47]. Plutchik's model is not the only tool used to assess emotional reactions. The Geneva emotional wheel (GEW) uses a circular structure with the axes defined by valence (bipolar subjective evaluation of positive/negative) and control to arrange 40 emotion terms in 20 emotion families [48]. The dimensional models organize emotions within a space with two or three dimensions along which the responses vary. Russell's Circumplex Model of Affect [49] encompasses valence on the x-axis, indicating the positive or negative component of emotion and arousal along the y-axis, reflecting the degree of mental activation or alertness that is elicited [50]. Arousal ranges from inactive (not excited) to active (excited, alert) [43]. Besides valence and arousal, a third dimension, called dominance, specifies the degree of control the subject exerts over the stimulus. Dominance ranges from a weak, helplessness feeling to a strong, empowered one. For instance, fear is defined as having low valence, high arousal and low dominance. From the behavioral decision-making perspective, we mention the approach-withdrawal (or appetitive-aversive) motivational model which reflects the tendency of approaching or rejecting the stimulus. According to [51], fear is generated by an aversive response that conducts to either active or passive physical reactions.

3.2. Physiological Data

Electroencephalography (EEG) non-invasively measures electrical potentials produced by neural activity which falls in the frequency ranges corresponding to the delta (<3 Hz), theta (3 Hz–8 Hz), alpha (8 Hz–12 Hz), beta (12 Hz–30 Hz) and gamma (>30 Hz) waves. EEG offers high precision time measurements—it can detect brain activity at a resolution of one millisecond—but unfortunately lacks spatial resolution. The recording area of an electrode is approximately one centimeter of the scalp, which corresponds to hundreds of thousands of neurons in the cerebral cortex. Thus, it is difficult to accurately pinpoint the exact source of brain activity or to distinguish between activities occurring at contiguous locations [52]. Moreover, EEG signals are prone to electrical interferences or artefacts resulting from body movements (eye blinks, muscular or cardiac activity) or environmental causes.

The right hemisphere processes negative emotions or aversive behaviors, while the left one is involved in mediating positive emotions or approach behaviors. The people who experience negative feelings, who are angry, afraid or depressed, present activations in the amygdala (part of the limbic forebrain) and in the right prefrontal cortex [53]. The literature largely supports the approach/withdrawal model of alpha asymmetry, which states that activation in the right cortical area (low alpha waves) is associated with an aversive behavior, while activation in the left cortical area indicates positive feelings [54–57].

Park et al. [58] observed an increase of the beta waves at the left temporal lobe when the users experienced fear. The work of [59] showed reduced beta power in the bilateral temporal and right frontal cortex for individuals suffering from panic disorders. An increase of beta intensity in the left temporal lobe was also noticed in [58] whenever the subjects felt threatened.

The research performed by [60] showed that the patients who experienced fear exhibited high theta, delta and alpha absolute power and low beta levels. The authors suggest that the increase of the alpha waves accompanies and regulates the excessive excitation of the slow waves in the temporal regions and in the limbic system. In [61], a patient suffering from agoraphobia and panic attacks had an increase in the beta activity and a sudden decrease of frontal-central theta power. Time-domain EEG analysis indicated a reduced P300 Event Related Potential (ERP) and an increase in the beta activity in the right temporal lobe, an increase in the alpha activity in F4 and a decrease of the T5 theta activity [62].

In [63], a negative relationship was observed between delta and alpha 2 activity. A decreased beta-delta coherence in anxious individuals was shown in [61], together with a significant decrease in delta during panic attacks. Beta activity in the central part of the frontal cortex increased, being accompanied by a significant reduction of the theta waves all over the cortex, similar to what has been found in [64].

The ratio of slow waves to fast waves (SW/FW) has a negative correlation with fear [65–67]. There was a statistically significant reduction in the SW/FW ratio (delta/beta and theta/beta) in the left frontal lobe in an experiment where data has been recorded from a single electrode [68]. Neutral states are reflected in equal levels of activation in both hemispheres [69]. Quantitative EEG studies, and in particular coherence (linear synchronization between EEG signals measured at different brain locations), indicated a lower degree of inter-hemispheric functional connectivity at the frontal region and intra-hemispheric at the temporal region [70].

Plethysmography (PPG) is a non-invasive circulatory assessment method that uses an infrared photoelectric sensor to record changes in blood flow from the finger or from the ear lobe. It determines blood volume pulse by calculating how much of the emitted light is reflected back. The PPG values are converted into heart rate, which is measured in beats per minute (bpm). Heart rate variance is a strong indicator of emotion. In [71], a decrease in variance while the heart rate was high was an indicator of fear. Heart rate, combined with other variables, can successfully classify emotions [72,73], although in others it was found that it had the smallest contributing factor [74].

Electrodermal Activity (EDA) or Galvanic Skin Response (GSR) is a measure of sweat glands production and therefore skin activity, in direct relation with the sympathetic nerve's state of excitation. GSR has two main components: tonic skin conductance, the baseline value recorded when no emotional stimulus is applied and phasic skin conductance, the response acquired when environmental and behavioral changes occur [50]. Increased GSR indicates arousal. It was the main contributing factor for emotion classification in several studies, including [75,76], being effective for discriminating fear from other negative emotions [77]. GSR recording devices are comfortable for users due to their light, easily attachable sensors [78].

In conclusion, we consider that the most relevant physiological signals to account for in fear assessment experiments are GSR, HR and the values of the alpha, beta and theta waves. In addition, the ratio of slow/fast waves is a good indicator of fear, together with alpha asymmetry—the difference in cortical activation between the right and left hemisphere in the alpha frequency band.

4. Physiological Data in VR-Based Machine Learning Applications for Treating Phobias

Virtual Reality can induce the same level of anxiety as real-life situations, and physiological data can be used to reflect stress level. In this section, we perform a short review on physiological data analysis in VR and on the ML techniques involved in emotion recognition and phobias treatment.

4.1. Physiological Data in VR-Based Applications for Treating Phobias

In the study presented in [79], the authors investigated the physiological responses of both nonphobic and phobic subjects in the VR environment. They monitored the skin resistance (SR), heart rate (HR) and skin temperature of 36 participants suffering from fear of flying and 22 participants with no fear. The anxiety level of the phobic participants was evaluated using Subjective Units of Distress, on a scale from 0 to 100 (0—no anxiety, 100—highest anxiety). The results showed a significant difference in the case of SR between two groups and no major difference in the case of HR and skin temperature. More intensive VR-based therapy sessions applied on the phobic subjects had a greater effect on 33 persons who succeeded to fly by plane after the VR treatment.

More physiological data was recorded in the experiment performed in [80], which confirmed the following hypotheses: virtual heights increased the subjects' stress levels and the cognitive load during beam-walking was higher in VR. Heart rate variability, heart rate frequency power, heart rate, electrodermal activity and EEG data have been recorded and analyzed to validate the two hypotheses. Heart rate variability varied from 6.6 beats/min in the unaltered low view to 7 beats/min in low VR conditions and 8.3 beats/min in high VR conditions. Heart rate started from 92 beats/min in unaltered view, continued with 97 beats/min in VR low and 97.1 beats/min in VR high conditions.

Electrodermal activity of five subjects was analyzed in [81] to measure stress level in VR conditions. The participants have not been diagnosed with acrophobia, but they claimed a certain discomfort in height situations. Each subject underwent a 15 min session consisting of three sub-sessions: height exposure in the real world (standing on the balcony of a building); height exposure in VR (the users did not interact with the VR environment); and height exposure in VR with VR environment interaction. The results proved that interaction with the environment during phobia treatment is important and that physiological measurements help in assessing emotional states.

Human responses to fear of heights in immersive VR (IVR) conditions were investigated in [82]. The authors performed two experiments: the first experiment on 21 subjects with ages ranging from 20 to 32 years and the second on 13 subjects with ages in the interval 20–27 years. During the first experiment, in which the subjects were exposed to four heights: 2, 6, 10, and 14 m in IVR conditions, GSR, heart rate and the participants' view direction were measured. In the second experiment, the subjects were exposed till 40 m in an immersive virtual environment. The authors measured physiological responses and head motion. Also, the participants had to report the perceived anxiety level. The results showed that there was a correlation between the anxiety level and the subjects' head pitch angle and that the anxiety level is accurately visible in phasic skin conductance responses. Also, it was established a correlation between anxiety/height and GSR measurements.

4.2. Machine Learning for Emotion Recognition

Automatic emotion recognition has gained the attention of many researchers in the past few decades. As of now, there are three major approaches to automatic emotion recognition: the first approach consists in analyzing facial expressions and speech, the second approach uses the peripheral physiological signals, and the third approach uses the brain signals recorded from the central nervous system [83]. Certainly, a method that will embrace all these three approaches will provide the best results. The emotion recognition models are used in applications such as man-machine interfaces, brain-machine communications, computer-assisted learning, health, art, entertainment, telepresence, telemedicine and driving safety control [84–86].

Machine Learning offers computers the ability to learn from large data sets [87]. Among the ML techniques, Deep Learning is increasingly used in various applications, due to its higher accuracy when huge amounts of data are used for training. For emotion recognition, different ML techniques have been employed.

A research tool called the Multimodal Affective User Interface is proposed in [85] for emotion discrimination. To obtain an accurate and reliable recognition tool, the system's inputs were "physiological components (facial expressions, vocal intonation, skin temperature, galvanic skin response and heart rate) and subjective components (written or spoken language)" [85]. Using short films as stimuli for eliciting emotions and the GSR, temperature and heart rate records from 29 subjects, the authors implemented three ML algorithms: kNN, Discriminant Function Analysis (DFA) and Marquardt Backpropagation (MBP), in order to obtain six classes of emotions (sadness, anger, surprise, fear, frustration and amusement). The reported recognition accuracies were: kNN—67% for sadness, 67% for anger, 67% for surprise, 87% for fear, 72% for frustration and 70% for amusement; DFA—78% for sadness, 72% for anger, 71% for surprise, 83% for fear, 68% for frustration and 74% for amusement; MBP—92% for sadness, 88% for anger, 70% for surprise, 87% for fear, 82% for frustration and 83% for amusement. Also, the authors pointed out "that detection of emotional cues from physiological data must also be gathered in a natural environment rather than in one where emotions are artificially extracted from other naturally co-occurring states" [85].

A stack of three autoencoders with two softmax classifiers was used in the EEG-based emotion recognition system proposed in [86]. 230 power spectral features of EEG signals extracted in 5 frequency bands (theta, lower alpha, upper alpha, beta and gamma) and the differences between the spectral powers of all the 14 symmetrical pairs of electrodes on the right and on the left hemispheres have been used as inputs for some DL networks. The efficiency of the system was evaluated in four experimental setups: DLN-100 using a DL network with 100 hidden nodes on each layer; DLN-50 using a DL network with 50 hidden nodes; DLN-50 with PCA (Principal Component Analysis to address the overfitting problem); and DLN-50 with PCA and CSA (Covariate Shift Adaptation to solve the problem of non-stationarity in EEG signals). The accuracies obtained for each experiment were: DLN-100: 49.52% for valence and 46.03% for arousal; DLN-50: 47.87% for valence and 45.50% for arousal; DLN-50 with PCA: 50.88% for valence and 48.64% for arousal; DLN-50 with PCA and CSA: 53.42% for valence and 52.03% for arousal.

A comprehensive review of physiological signals-based emotion recognition techniques is presented in [88]. 16 studies including various classifiers such as Support Vector Machine, Linear Discriminant Analysis, k-Nearest Neighbors, Regression Tree, Bayesian Networks, Hidden Markov Model, Random Forest, Neural Networks, Canonical Correlation Analysis, Hybrid Linear Discriminant Analysis, Marquardt Back Propagation, Tabu search, and Fisher Linear Discriminant Analysis are compared with respect to their accuracies, bio-signal data, stimuli employed and feature extraction techniques. Emotions are considered in two models: discrete and dimensional. In the case of user dependent systems, the best performance (accuracy 95%) was achieved using linear discriminate in a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) for joy, anger, sad and pleasure classification [89]. The bio-signals used were: and Electromyogram, Electrocardiogram, Skin Conductance, Respiration. An accuracy of 86% was obtained to classify joy and sadness in the case of user independent system [90]. The ECG feature extraction was performed using a non-linear transformation of the first derivative and tabu search was involved to acquire the best combination of the ECG features.

Bayesian classifiers are used in a multimodal framework for analysis and emotion recognition [91]. Eight emotional states: anger, despair, interest, pleasure, sadness, irritation, joy and pride were recognized based on facial expressions, gestures and speech. The authors reported that all emotions except despair can be recognized with more than 70% accuracy and the highest accuracy was recorded for anger recognition (90%) [91].

A Deep Convolutional Neural Network-based approach for expression classification on the EmotiW (The Emotion Recognition in the Wild contest) dataset is presented in [92]. Seven basic expressions (neutral, happy, surprised, fearful, angry, sad and disgusted) were recognized, with an overall accuracy of 48.5% in the validation set and 55.6% in the test set.

The usage of VR environments as stimuli for human emotion recognition has barely been studied. In most research regarding automatic recognition of human emotions, the stimuli were either images, sequences of films or music. One of the first reports on EEG-based human emotion detection using VR stimuli is presented in [93]. Four deep neural networks were tested: standard, deep network with dropout, deep network with L1 regularization and deep network with dropout and L1 regularization. The last one achieved a 79.76% accuracy. Also, a high classification accuracy, close to 96%, was obtained for excitement detection while being immersed in a VR environment.

In [11], the physiological data of 20 Tetris players were recorded and analyzed using three classifiers: LDA, Quadratic Discriminant Analysis (QDA) and SVM. The results showed that playing the Tetris game at different levels of difficulty gives rise to different emotional states. Without feature selection, the best classifiers obtained an accuracy of 55% for peripheral signals and 48% for EEG (LDA, followed by SVM). Feature selection increased the classification accuracy to 59%, respectively, 56%. After the fusion of the two signal categories, the accuracy increased to 63%. A comparative study of four popular ML techniques aimed at identifying the affective states (anxiety, engagement, boredom, frustration and anger) of users solving anagrams or playing Pong is presented in [94]. The authors reported that SVM with a classification accuracy of 85.81% performed the best, closely followed by RT (83.5%), kNN (75.16%) and Bayesian Network (74.03%) [94]. A Dynamic Difficulty Adjustment (DDA) of game levels based on physiological data is presented in [12]. The authors used psychological responses during gameplay and a RT-based model for recognizing anxiety levels (low, medium, high). The model gave 78% correct predictions [12]. However, the adjustment was based on clauses and conditions, not on a prediction method.

A more detailed investigation of ML techniques used in emotions classification was performed in [95].

Fourteen physiological signals were recorded in VR conditions and used for emotion recognition in [96]: EEG f4, vertical and horizontal Electrooculography (EOG), Electromyography (EMG), Electrodermal Activity (EDA), Electrocardiogram (ECG), Chest Respiration (RIP), Abdomen Respiration (RIP), Peripheral Temperature, Heart Rate via PulseOx, Blood Volume (PPG) via PulseOx, Blood Oxygen (SpO2) via PulseOx, Head Acceleration and Rotation, Body Acceleration and Rotation. The Naive Bayes, k-Nearest Neighbor and Support Vector Machine techniques have been used to perform a binary classification: high-arousal or moderate/low arousal. The best accuracy was achieved in the case of SVM (89.19%).

4.3. Machine Learning for Identifying Anxiety Level in Phobia Therapy

In [97], a deep convolutional network was used to detect acrophobia level (level 1 = only somewhat strong or not strong, level 2 = moderately strong, level 3 = quite strong, level 4 = very strong). However, a tailored treatment was not performed. Richie's Plank Experience was used as the virtual environment, and EEG data from 60 subjects was acquired to feed a deep learning network model VGG-16. The performance of the model has been measured using the accuracy, recall and precision parameters. The average accuracy obtained was 88.77%.

A VRET system used to overcome public speaking anxiety, fear of heights and panic disorder is described in [98]. The system contains a mental stress prediction framework, which uses data extracted from GSR, blood volume pressure (BVP) and skin temperature signals to predict anxiety level. 30 persons participated in the experiments from [98], focused on public speaking anxiety. Four classes were defined for anxiety level: low, mild, moderate and high, and a SVM classifier with radial basis function (RBF) as kernel was used to train the models with various window lengths: 3, 5, 8, 10, 13, 15, 18, 20, 23, 25, 28, 29, and 30 s. A comparison between models was performed, and the results

highlighted that the model using signal fusion outperformed the models using standalone signals. The early fusion method achieved the best accuracy of 86.3%. Model training and data processing were not performed during the experiments (Table 1).

Table 1. Performance in phobia level classification using ML.

	Classifiers	Goal	Signals	Number of Subjects	Performance or Significant Results
[97] 2018	CNN with VGG-16	Detect acrophobia level	EEG	60 subjects	average accuracy 88.77%
[98] 2019	SVM with RBF kernel	Predict anxiety level (public speaking fear)	GSR, BVP, skin temperature	30 persons	BVP accuracy window size 18 s 74.1% GSR accuracy window size 23 s 76.6% Skin temperature accuracy window size 18 s 75.1% Signal fusion (early) window size 20 s 86.3% Signal fusion (late) window size 20 s 83.2%

Currently, VRET is seen as an efficient method for phobia treatment, both from a financial and a comfort point of view. It offers flexibility, confidentiality and trust, encouraging more people to seek treatment [16,96].

As far as we know, the issue of classifying emotion levels in VR conditions, meaning how intensely an emotion is felt based on different factors, has not been yet properly defined.

In the proposed system, we focus our study on the ML and DL methods, which automatically classify fear level using physiological data. The dataset has been acquired in direct relation to our acrophobia therapy application, more specifically, by exposing the users to different heights in both the real-world and virtual environment.

5. The Machine Learning and Deep Neural Networks Approach for the Acrophobia VRET Game

The proposed VR system contains an ML-based decision support that adjusts the playing scenario according to the patient's level of fear. It incorporates a real-time decision engine which uses the patient's physiological data and determines the game level to be played next. In our ML-based decision support, the data obtained from the users contribute to configuring the game in order to suit each patient's individual characteristics.

For this purpose, two classifiers were used: one to estimate the patient's current fear level (C1) and one that determines the appropriate treatment according to the target fear level (C2). In our previous approaches [99,100], we used only deep neural networks as classifiers, but the obtained results pushed us to continue to test with various ML techniques. In this paper, we extended our work by defining a ML-based decision support that relies on various ML techniques such as SVM, kNN, LDA, RF and 4 deep neural network models (Figure 1).

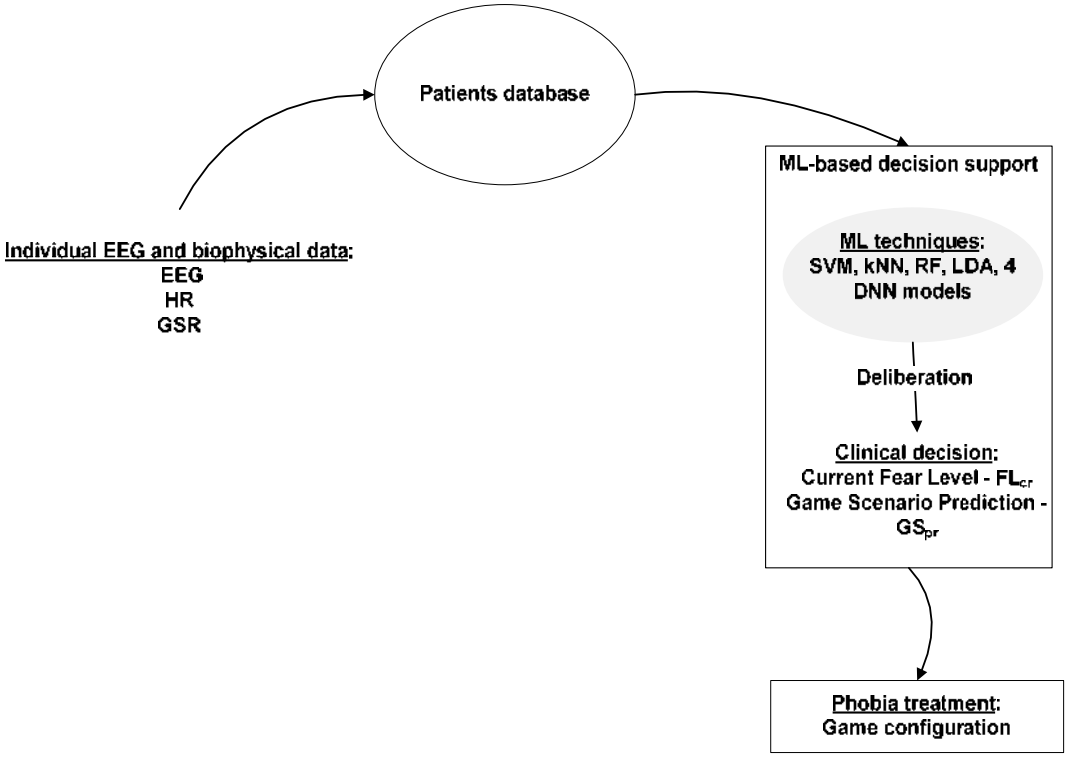

Figure 1. ML-based decision support for phobia treatment.

As in our previous work, we used two different fear level scales [99,100]:

- 2-choice scale, with 2 possible values, 0 and 1. 0 stands for relaxation and 1 stands for fear.
- 4-choice scale, with 4 possible values (0–3). 0—complete relaxation, 1—low fear, 2—moderate fear and 3—high level of anxiety.

The game scenarios consist of different game levels, each game level corresponding to a certain degree of height exposure in different contexts or a combination of certain height exposure degrees. The data recorded in real time from the patient is fed to the C1 classifier and the current fear level (FL_{cr}) is computed. C1 estimates the level of fear the patient currently experiences.

To determine the target fear level (FL_t) that ensures a gradual and appropriate exposure to height, we used the following formulas:

2-choice scale	**4-choice scale**
if FLcr = = 0 then FLt = 1	if FLcr = = 0 or FLcr = = 1 then FLt = FLcr + 1
if FLcr = =1 then FLt = 0	if FLcr = = 2 then FLt = FLcr
	if FLcr= = 3 then FLt = FLcr − 1

The target fear level (FL_t), together with the patient's physiological data, are fed to the C2 classifier and the next game scenario (GS_{pr}) is predicted. C2 estimates the phobia treatment.

The user plays the predicted level of the game and new physiological data is acquired. C1 computes a new general fear level and C2 predicts the game scenario to be played next. The process goes on for as long as the patient or the therapist consider appropriate—the patient can exit the game at any time if he/she feels uncomfortable—or a total predefined number of scenarios is reached.

6. Experimental Methodology

The experiment was conducted in summer–autumn 2018 and involved the participation of 8 subjects who played an acrophobia game while their physiological (HR and GSR) and EEG data were recorded. The experiment was approved by the ethics committee of the UEFISCDI project 1/2018 and

UPB CRC Research Grant 2017 and University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers. Prior to the experiment, the subjects signed a consent form and filled in a demographic and a Visual Height Intolerance questionnaire [101]. Prior to the tests, they were informed about the purpose of the experiment and research objectives. Moreover, they were presented with the steps of the procedure and the experimenter made sure that they fully understood what they were required to do. From the 8 users (aged 22–50 years, 6 women and 2 men), 2 suffered from a mild form of acrophobia, 4 from a medium-intensity fear of heights and 2 experienced a severe form of height intolerance. This classification resulted by assessing the responses to the Visual Height Intolerance questionnaire. More details can be found in [99,100]. They did not consume coffee or other energizing drinks before the experiment and made sure they had a relaxing sleep in the previous night. With respect to the therapy history, our subjects have not undergone any phobia alleviation treatment beforehand, neither medical nor psychological. Half of the users had previous experience in using VR systems and the others had not. For the second category, we provided some VR introductory sessions to accommodate them with the VR perception. Thus, we explained to them what a VR environment represents, which are the hardware components (VR glasses, controllers, sensors), how they work and how they can be adjusted. We presented the users the actions occurring in the game when each of the buttons from the controllers are pressed. Then, the subjects played a basic demo game which accommodated them with the VR perception.

The EEG data have been acquired using the Acticap Xpress Bundle [102] device with 16 dry electrodes, while HR and GSR have been recorded via Shimmers Multi-Sensory [103]. The next exposure scenario has been predicted in real-time by C2, based on the EEG, physiological data and the target fear level. The target fear level was calculated according to the formulas mentioned above, by taking into account the patient's current fear level. The current fear level was estimated by C1.

The classifiers we used were: kNN, SVM with linear kernel, RF, LDA and 4 deep neural network models with a varying number of hidden layers and neurons per layer.

6.1. Experiments and Dataset Construction

The 16 dry electrodes of the Acticap Xpress Bundle device [102] were placed according to 10/20 system in the following locations: FP1, FP2, FC5, FC1, FC2, FC6, T7, C3, C4, T8, P3, P1, P2, P4, O1 and O2. The log-normalized powers of all the 16 channels in the alpha, beta and theta frequency ranges were recorded and pre-processed in real-time for artefact removal. The ground and reference electrodes were attached to the ears. Using the Shimmer3 GSR+ Unit [104] of the Shimmers Multi-Sensory device, we acquired the subjects' electrodermal activity and heart rate values. The Shimmer3 GSR+ Unit, which has Bluetooth connectivity, measures the skin's electrical characteristics in microSiemens and captures a PPG signal (using the Shimmer optical pulse probe) that is later converted to estimate heart rate (HR).

The two classifiers C1 and C2 have been fed with training data originating from two preliminary experiments where the subjects have been both in vivo and virtual y exposed to the first, fourth and sixth floors of a building, as well as on the ground floor, at 4 m, 2 m and a few centimeters away from a terrace's railing. In the virtual environment, the players have been also exposed to the view from the building's rooftop. The experiment in the virtual environment (consisting of three sessions, expanded over three days) has been preceded and succeeded by a real-world session. The EEG, GSR and HR data has been collected, together with the user's perceived level of fear, called Subjective Unit of Distress (SUD), during each trial. Each patient was required to rate his/her fear on the 11-choice scale, a gradual scale with values from 0 to 10, where 0 corresponds to complete relaxation and 10 to extreme fear. The modality of reporting the SUD was verbally for the in vivo experiment and by pointing a virtual laser with the controller on a panel in the virtual environment (Figure 2).

The acrophobia game was rendered on the HTC Vive head-mounted display [33]. Interaction in the virtual environment was ensured through the controllers, so that the player advances in the game by teleportation—he/she presses on the floor in the virtual environment at various positions where

he/she wants to go, is free to navigate wherever he/she wants, but he/she has to accomplishes the tasks of collecting coins of different colors (bronze, silver and gold) at 4 m, 2 m and 0 m distance from the balcony's railing at ground level and at the first, fourth, sixth floors, as well as on the roof of the building. A coin is collected by bending and grabbing it with the controller. The game contained only visual and vestibular stimuli. There were no audio cues or animations to accompany the graphical presentation.

In both the real-world and virtual environment, each user totalized several 63 trials (3 sessions × 5 building levels × 3 distances from the railing = 45 in the virtual environment and 2 sessions × 3 building levels × 3 distances from the railing = 18 in the real-world environment). We thus obtained a dataset of 25,000 entries on average for each patient, which was saved in a database and used for training classifiers C1 and C2.

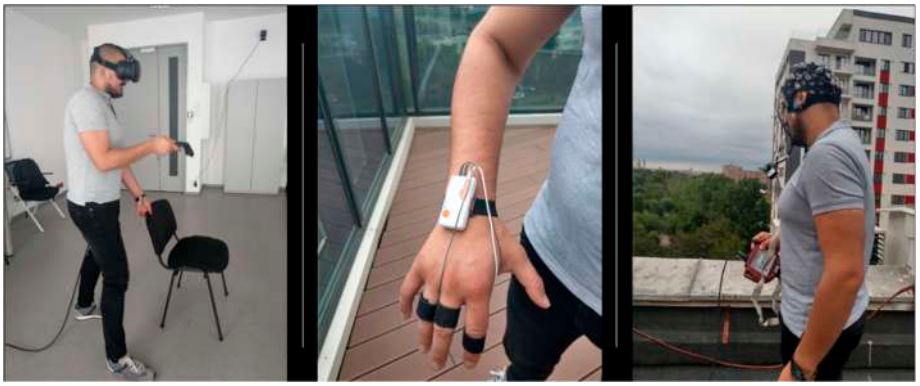

Figure 2. User during in vivo and virtual exposure with physiological signals monitoring.

For training C1, we had as input features the physiological data recorded during the 63 trials—the EEG log-normalized powers of the 16 channels in the alpha, beta and theta frequencies, the GSR and HR values. The output feature was the fear level (SUD) on three scales: the 11-choice scale (values as they were recorded, ranked from 0 to 10), 4-choice (fear rates from 0 to 3) and 2-choice (values of either 0 or 1).

The ratings from the 11-choice-scale have been grouped into 4 clusters in order to create the 4-choice-scale (Table 2):

- 0 (relaxation)—rating 0 in the 11-choice-scale
- 1 (low fear)—ratings 1–3 in the 11-choice-scale
- 2 (medium fear)—ratings 4–7 in the 11-choice scale
- 3 (high fear)—ratings 8–10 in the 11-choice scale

Similarly, the ratings from the 4-choice-scale have been grouped into 2 clusters in order to create the 2-choice scale:

- 0 (relaxation)—ratings 0–1 in the 4-choice scale
- 1 (fear)—ratings 2–3 in the 4-choice scale.

The classifiers we used were: kNN, SVM with linear kernel, RF, LDA and 4 deep neural network models with different numbers of hidden layers and neurons per hidden layer. We have chosen these classifiers because they have been widely used in the literature (see Sections 4.2 and 4.3). SVM provides the best results for emotion classification. kNN is used for signal classification. LDA has been used for binary and multi-class classification, being highly employed in the medical field. RF is a top classifier and the deep neural networks provide good classification results due to their ability to learn high-level features from large amounts of data in an incremental way.

Table 2. Fear level classification scales.

11-Choice-Scale	4-Choice-Scale	2-Choice-Scale
0	0 (relaxation)	0 (relaxation)
1	1 (low fear)	
2		
3		
4	2 (medium fear)	1 (fear)
5		
6		
7		
8	3 (high fear)	
9		
10		

The Sequential Forward Selection (SFS) feature selection algorithm was applied for kNN, RF and LDA. kNN is a non-parametric, feature similarity-based method used especially for classifying signals and images. The decision is made by taking into account the class of the majority of the k-nearest neighbors. SVM is a supervised machine learning algorithm that finds the hyperplane best segregating two or more classes. RF operates by constructing an ensemble of decision trees. The predicted class is obtained by combining the prediction of all individual trees, based on the "bagging" method stating that a combination of learning models increases the overall result. LDA is a dimensionality reduction technique that projects the dataset onto a lower dimensional space and finds the axes that maximize the separation between multiple classes, avoiding overfitting and reducing the computational cost. All these algorithms were run in Python, using their corresponding implementations from the scikit-learn library [105].

Using the TensorFlow deep learning framework [106], we created four Keras [107] sequential models for binary and multi-class classification: DNN_Model_1, DNN_Model_2, DNN_Model_3 and DNN_Model_4 (Table 3). Each network has an input layer of 50 neurons (16 neurons for the alpha values, 16 for the beta values, 16 for the theta values, 1 for GSR and 1 for HR) and an output of one neuron, corresponding to the predicted level of fear. Before training, the data has been standardized to reduce it to zero mean and unit variance. We performed a 10-fold cross-validation procedure 10 times and saved the weights of each network in .hdf5 files, together with the corresponding accuracies. The 10-fold cross-validation procedure was computed using the functionalities implemented in the scikit-learn library for k-fold cross-validation, with k = 10. The procedure has one parameter k that represents the number of groups the data is split into. Each group is taken as a test data set and the remaining k − 1 groups are taken as training data set. Then, the model is fit on the training set and tested on the test set. The evaluation score is retained, and the model is discarded. In the end, the cross-validation accuracy is calculated based on the k evaluation scores computed at each step.

Finally, the model version with the highest accuracy for each network has been selected and further used in the experiment. This procedure was repeated for every user, for the 2-choice, 4-choice and 11-choice scales. This technique was applied and published in [99,100]. In the current stage of research, we also trained ML classifiers (kNN, RF, LDA and SVM) in the same way—for every user, 10 times, for each fear scale—and the model providing the highest accuracy was saved for further use.

Table 3. Properties of the Deep Neural Network models.

DNN Models	Activation Function	Activation Function in the Output Layer	Loss Function	Optimization Algorithm	Epochs and Batch Size
DNN_Model_1 3 hidden layers, with 150 neurons on each hidden layer	Rectified Linear Unit (RELU)	2-choice scale Sigmoid activation function	2-choice scale Binary crossentropy	Adam gradient descent	1000 epochs for training
DNN_Model_2 3 hidden layers, with 300 neurons on each hidden layer					
DNN_Model_3 6 hidden layers, with 150 neurons on each hidden layer		4-choice scale Softmax activation function	4-choice scale Categorical crossentropy and one-hot encoding		Batch size of 20
6 hidden layers, with 300 neurons on each hidden layer					

Classifier C2 predicts the game level that should be played next, i.e., the next exposure scenario (parameter GS_{pr}). The Sequential Forward Selection (SFS) feature selection algorithm has been applied for kNN, RF and LDA. For training C2, the deep learning and machine learning models received as inputs the EEG, GSR, HR and SUD values, while the output represented an encoding of the height where these physiological values have been recorded—0 for ground floor, 1 for the first floor, 2 for the fourth floor, 3 for the sixth floor, and 4 for the roof of the building. For testing classifier C2, we provide as input EEG, GSR, HR and target fear level (FLt) and obtain as output the encoding of the height where the player should be taken to in the game (from 0 to 4, as mentioned above). Thus, if the user is currently feeling anxious (FLcr = 3), we calculate a target fear level FLt = 2 (so we want him to feel less anxious) and feed this value as input to classifier C2 in order to generate for us the next exposure scenario GSpr, on a scale from 0 to 4: 0 for ground floor, 1 for the first floor, 2 for the fourth floor, 3 for the sixth floor and 4 for the roof of the building.

The same DNN models were used for classifier C1, with the same number of hidden layers and neurons on each hidden layer (Table 3). Each network had an input layer of 51 neurons (16 neurons for the alpha values, 16 for the beta values, 16 for the theta values, 1 for GSR, 1 for HR and 1 for the "target fear level" feature). The output represented the level in the building from where the user should restart playing the game. The method for obtaining a personalized height exposure model to be validated on test dataset was: we repeated the 10-fold cross-validation procedure 10 times for each subject and saved the weights and the corresponding accuracies of each network in .hdf5 files; the model version with the highest accuracy for each network has been selected and further used in the experiment for all fear scales.

The ML classifiers (kNN, RF, LDA and SVM) were trained in the same way—for every user, 10 times, for each fear scale—and the model resulting in the highest accuracy was saved for further use. For cross-validation, the data has been divided into 70% training and 30% test.

For computing the accuracy of the classifiers, both a user-dependent and a user-independent modality were used. Each classifier was trained using the data of the other subjects in the user-independent modality. We applied the trained model on the data of the tested user in order to calculate cross-validation and test accuracies. This approach makes it possible to calculate the performance of the classifiers in the worst possible case, where the model lacks user specificity. On the other hand, in the case of the user-dependent modality, for each subject, each classifier has been trained, cross-validated and tested on his/her own data. Feature selection has been also computed for each

classifier in order to improve generalization across subjects. We used Sequential Forward Selection (SFS), a greedy algorithm that reduces the d-dimensional space to a k-dimensional space. In our case, we set k to 20, so that it would extract the most relevant 20 features from the total number of 50 features (16 EEG channels for the alpha, beta and theta waves, GSR and HR). The goal of feature selection was two-fold: we wanted to improve the computational efficiency and to reduce the generalization error of the model by removing irrelevant features or noise. SFS has been applied for kNN, RF and LDA.

6.2. The Acrophobia Game

The the game, which has been developed using the Unity engine [108], was synchronized in real-time with the Open Vibe [109] application for collecting EEG signals and with the Shimmer3 GSR+ Unit that records GSR and HR via Lab Stream Layer (LSL) [110]. Using a multi-thread C# application, we ran 5 threads simultaneously: one for recording the input from the game (fear ratings, events from the game, such as when the coins have been collected or when a level has been finished), peripheral physiological data (HR and GSR from the Shimmers3 Unit), alpha, beta and theta power spectral densities. At each session, 5 separate log .csv files were generated, each of them containing the timestamps (a timestamps represents the number of milliseconds passed since 1st January 1970) and the recorded data (either from the game, peripheral, alpha, beta or theta). The EEG data is extracted at an interval of 62.5 ms and the GSR and HR values were extracted at an interval of 19.53 ms. As the data has been saved at different sampling frequencies, we developed another processing module that merged the information from the log .csv files, averaged and aligned them according to the timestamps in order to have a compact dataset of EEG and peripheral recordings mapped onto the events occurring in the game.

In order to extract the EEG data, we applied a bandpass Butterworth temporal filter, time-based epoching with the epoch duration of 1 s, and then squared the input values using the Simple DSP box from Open Vibe Designer. In addition, we averaged the signal and applied log-normalization using again the Simple DSP box. After all this preprocessing of the raw data, the alpha, beta and theta frequency powers have been extracted (Figure 3).

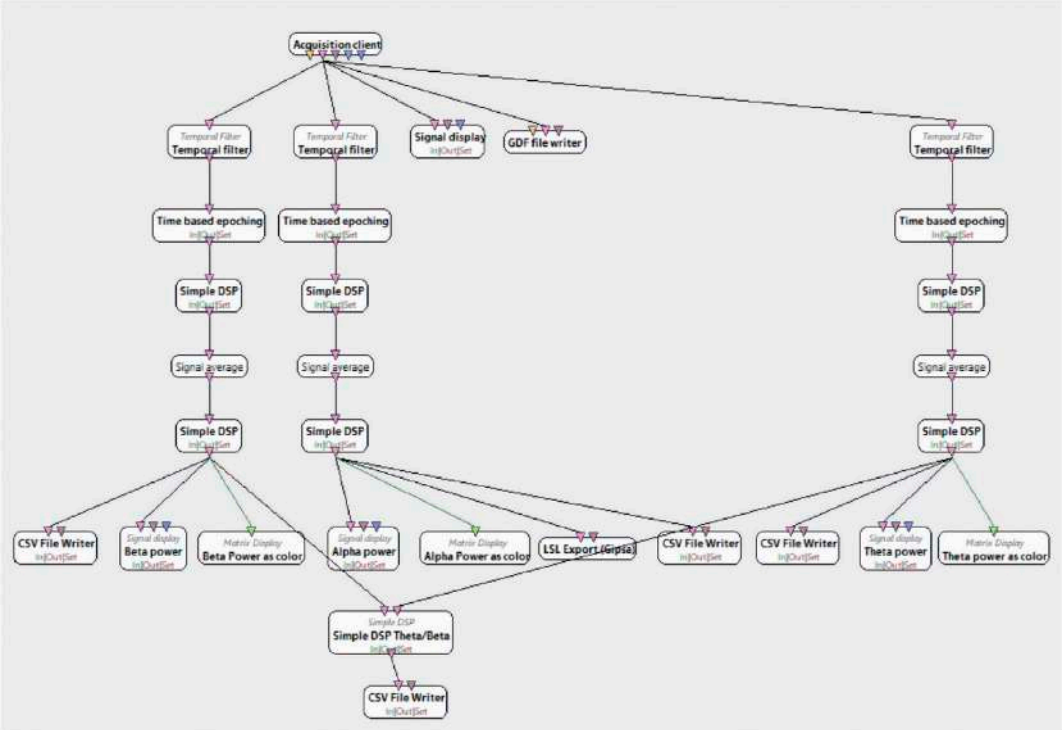

Figure 3. EEG signal recording and decomposition in frequency bands.

All data was denoised and preprocessed in real time by applying a method named "the last correct value", introduced by us. In our preprocessing module, all EEG and physiological data were inspected in real-time. As LSL was pulling sample data from the recording devices, before saving it into the corresponding log files, it was inspected to see if faulty values occur. For instance, if a negative value or one exceeding one and a half than the average of the previous values on a 5-seconds time span appeared, it was replaced with the average of the data recorded in the previous 5 s. If the device malfunctioned since the moment it started recording (suppose it took a longer time to initialize or calibrate), we initialized the last correct value with some average values—4.5 microVolts^2 for the EEG log-normalized powers, 1 microSiemens for GSR and 75 bpm for HR. This method has been applied because we could not manually remove the noisy data nor stop the recording whenever such type of artefacts occurred in real-time.

In addition, it was saved in a database in both processed and unprocessed format for ulterior study and analysis. At the start of the game, the user was placed on the ground floor, where he/she had to navigate freely in the scene and collect a bronze, a silver and a gold coin (Figures 4 and 5). The Shimmers Unit has been attached to the left hand and the right hand has been used for holding the HTC Vive controller. In this way, we tried to reduce the chances of introducing hand movement artefacts in the GSR and HR signals. At all time, the users were required to sit on a chair and move only their head and the right hand.

Figure 4. The virtual environment, view of the building from the ground floor.

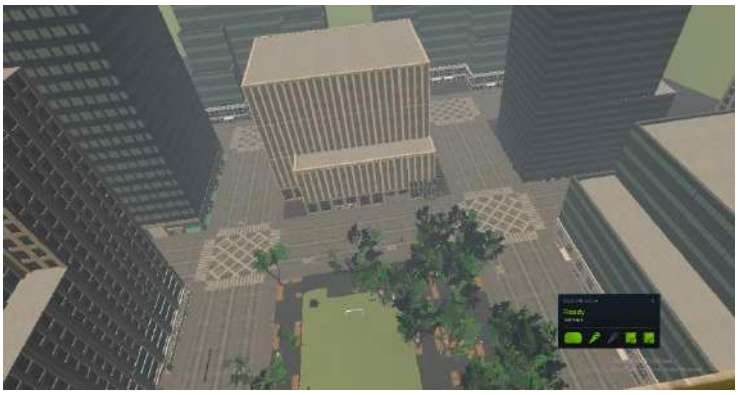

Figure 5. The view from the building's rooftop.

Consequently, he/she reported the perceived SUD by pointing with a virtual laser on a panel which contained a range of options from 0 to 10 for fear level evaluation. The physiological data were averaged and classifier C1 predicted the subject's current level of fear. To validate the accuracy of C1, we collected self-estimated SUDs. C1 predicted the current fear level based on the classification model created using the data from the previous experiment and a measure of assessing its accuracy was by comparing its output with the SUD perceived and acknowledged by the users directly during

gameplay (during each trial of the game, we also asked the users to report the perceived fear level (SUD)). This parameter was called test accuracy. Based on the EEG, physiological data and target fear level (obtained using the fear level estimated by classifier C1), the next level of exposure was determined by classifier C2, either on the 2-choice or on the 4-choice scale.

7. Results

The subjects played the game twice—once using the 2-choice and once using the 4-choice model. During each session, they were exposed to and played 10 scenes. At any time, the users could interrupt the game if they felt uncomfortable or the experimenter could terminate the session whenever he/she observed any abnormal events occurring. However, this was not the case, as all the subjects succeeded in completing both sessions without any difficulties. The maximum cross-validation accuracies on the training dataset and the validation (test) accuracies for each model, for both the player-independent and player-dependent modalities, with and without feature selection, are presented in Tables 4–7. The Test column for the C2 classifier is empty because we did not use any method for testing the accuracy of C2. This classifier has only been cross-validated on the training dataset.

With SFS feature selection, the most selected features were: for the 2-choice scale—alpha FC2, C3, T8, O2, beta P4, theta C3, T8 and HR; for the 4-choice scale—alpha FC5, C3, T8, P4 and O2, beta FP2, FC5, P4 and theta T8; for the 11-choice scale—alpha FC2, C3, T8, beta FP2, C5 and HR. We observed that the most important features where the alpha values in the right pre-frontal area, left central and right temporal, beta values in the frontal and parietal areas, theta values in the temporal area and the heart rate.

Table 4. Maximum cross-validation accuracy and test (validation) accuracy (in %) for the player-independent modality, without SFS feature selection.

Classifier Type	C1				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
SVM	80.5	64.75	60.5	46	59.5
kNN	99.5	43.75	99	52.75	98.25
RF	99.25	66.5	99	39.25	99
LDA	79.5	64.75	57.5	37.75	49.25
DNN_Model_1	95	58.3	87.825	45.425	79.4
DNN_Model_2	95.77	58.15	90.525	20.8	84.95
DNN_Model_3	94.75	58.3	86.55	37.7	74.025
DNN_Model_4	94.7	79.12	88.275	37.1	80.85
	C2				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
SVM	64.25	-	69	-	71
kNN	22.75	-	22.75	-	22.75
RF	99.75	-	100	-	100
LDA	24.5	-	25.75	-	29.5
DNN_Model_1	98.325	-	98.6	-	98.475
DNN_Model_2	98.5	-	98.725	-	98.3
DNN_Model_3	97.675	-	97.825	-	98.325
DNN_Model_4	97.8	-	98.15	-	97.575

Table 5. Maximum cross-validation accuracy and test (validation) accuracy (in %) for the player-independent modality, with SFS feature selection.

Classifier Type	C1				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
kNN	54	49.9175	32.25	30.24	25
RF	54.5	60.4175	33.25	38.5725	29.75
LDA	65.75	64.585	35.25	33.5725	25.25
	C2				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
kNN	32.75	-	36	-	41.75
RF	35.5	-	40.5	-	41.75
LDA	37.25	-	42.75	-	44.5

Table 6. Maximum cross-validation accuracy and test (validation) accuracy (in %) for the player-dependent modality, without SFS feature selection.

Classifier Type	C1				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
SVM	88	89.5	74.75	42.5	77.75
kNN	99.5	77	99	29.25	98.25
RF	99.75	77	99.25	21	99
LDA	87	60.5	71.25	21.75	64
DNN_Model_1	95.03	72.9	87.945	41.8975	79.485
DNN_Model_2	95.51	68.735	90.4975	24.9925	85.095
DNN_Model_3	94.4375	62.45	86.325	34.15	74.275
DNN_Model_4	94.575	54.125	88.28	38.325	80.45
	C2				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
SVM	82.75	-	86.5	-	86.5
kNN	23.75	-	23.75	-	23.75
RF	99.75	-	99.75	-	100
LDA	23	-	20.5	-	27.5
DNN_Model_1	98.4	-	98.675	-	98.75
DNN_Model_2	98.725	-	98.5	-	98.65
DNN_Model_3	97.45	-	97.825	-	98.5
DNN_Model_4	97.375	-	97.775	-	98.175

Table 7. Maximum cross-validation accuracy and test (validation) accuracy (in %) for the player-dependent modality, with SFS feature selection.

Classifier Type	C1				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
kNN	76.75	72.9175	52.25	16.665	42
RF	77	68.75	49.75	28.5725	45.75
LDA	81	85.4175	54.5	17.5	40.5
	C2				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
kNN	50.25	-	52.25	-	53.25
RF	50.5	-	53.5	-	56.5
LDA	52	-	56	-	56.75

The RF algorithm adds the benefit of computing the relative importance of each feature on the prediction. The implementation in the scikit-learn library measures feature importance by looking at how much the tree nodes using that feature reduce impurity for all the trees in the forest. Table 8 presents the most relevant 15 features, in descending order according to their importance, for the 2-choice, 4-choice and 11-choice scales, for both classifiers, for the player-independent modality. Table 9 contains the same attributes, but for the player-dependent modality. FL_t stands for "target fear level", B_ for "beta", A_ for "alpha" and T_ for "theta". Thus, B_C3 represents the beta value of the C3 electrode (central scalp position, left side). A_FC6 represents the alpha value of the FC6 electrode (fronto-central position, right side).

Table 8. Feature (F) and feature importance (FI) for the player-independent modality.

C1						C2					
2-Choice Scale		4-Choice Scale		11-Choice Scale		2-Choice Scale		4-Choice Scale		11-Choice Scale	
F	FI	F	FI	F	FI	F	FI	F	FI	F	FI
GSR	0.41	GSR	0.45	GSR	0.49	GSR	0.44	FL_t	0.69	FL_t	0.87
HR	0.28	HR	0.28	HR	0.24	FL_t	0.37	GSR	0.41	GSR	0.39
B_C3	0.15	B_FC6	0.15	B_FC6	0.14	HR	0.23	HR	0.20	HR	0.18
B_P3	0.13	B_C3	0.13	B_FC5	0.12	B_FC6	0.14	A_FC6	0.12	B_FC6	0.13
B_FC2	0.13	B_FC2	0.12	B_C3	0.12	A_FC6	0.13	B_FC6	0.12	A_FC6	0.11
B_FC6	0.13	B_FP1	0.12	B_FC2	0.12	B_FC5	0.10	B_P3	0.10	B_P3	0.09
B_FP2	0.12	B_P3	0.12	B_P3	0.11	T_FC6	0.10	B_T8	0.09	B_FC2	0.09
A_FC6	0.12	T_FC6	0.12	T_FC6	0.11	B_P3	0.09	B_FC2	0.09	T_FC6	0.08
B_C4	0.10	B_O1	0.11	B_FP1	0.10	B_T8	0.09	B_C3	0.09	B_T8	0.08
B_FC5	0.10	B_FC5	0.11	A_FC6	0.10	B_O1	0.09	T_FC6	0.08	B_FC5	0.07
B_FP1	0.09	B_T8	0.09	B_T8	0.10	B_C3	0.09	B_O2	0.08	B_O2	0.07
T_FC6	0.08	B_P2	0.09	B_O1	0.08	B_FC2	0.09	B_FC5	0.08	B_FP1	0.07
A_FP1	0.08	B_FC1	0.08	A_FP1	0.08	B_O2	0.09	B_FP1	0.07	B_C3	0.07
A_FP2	0.08	A_FP1	0.08	B_P2	0.08	B_P2	0.08	A_FP1	0.07	B_P2	0.06
B_T8	0.08	A_FC6	0.08	T_FP1	0.08	B_FP1	0.08	A_O1	0.06	B_O1	0.06

Table 9. Feature (F) and feature importance (FI) for the player-dependent modality.

C1						C2					
2-Choice Scale		**4-Choice Scale**		**11-Choice Scale**		**2-Choice Scale**		**4-Choice Scale**		**11-Choice Scale**	
F	FI	F	FI	F	FI	F	FI	F	FI	F	FI
GSR	0.40	GSR	0.46	GSR	0.48	GSR	0.54	FL_t	0.66	FL_t	0.79
HR	0.25	HR	0.32	HR	0.27	FL_t	0.32	GSR	0.47	GSR	0.42
B_FC2	0.22	B_FC6	0.17	B_FP1	0.14	HR	0.24	HR	0.20	HR	0.18
B_C4	0.15	B_FC2	0.16	A_FC6	0.14	A_FC6	0.15	B_FC6	0.14	T_FC6	0.12
B_FC6	0.14	B_P2	0.12	B_FC2	0.14	B_FC6	0.14	A_FC6	0.11	B_FC6	0.12
A_FP1	0.14	B_FP1	0.12	B_FC6	0.13	B_FP1	0.12	B_FC2	0.10	A_FC6	0.12
B_P2	0.13	T_FC6	0.11	T_FC6	0.12	T_FC6	0.10	T_FC6	0.09	B_P3	0.11
A_FC6	0.12	B_O1	0.10	B_O1	0.12	B_FC2	0.10	B_FC5	0.08	B_FC2	0.11
B_FP1	0.10	A_FC6	0.10	A_FP1	0.11	B_O2	0.09	B_O2	0.08	B_FP1	0.08
B_O2	0.10	A_FP1	0.10	B_FC5	0.11	B_P1	0.09	B_C4	0.08	A_FC1	0.08
T_P2	0.08	B_P3	0.10	B_P2	0.10	B_O1	0.08	B_FP1	0.07	T_FP1	0.07
T_FC6	0.08	B_C4	0.09	B_P3	0.10	A_O1	0.08	A_P4	0.07	A_O1	0.07
B_O1	0.08	B_FC5	0.09	B_C3	0.09	B_P2	0.08	A_FP1	0.07	B_T8	0.07
B_C3	0.08	A_P2	0.08	B_T8	0.09	T_P3	0.07	B_P2	0.07	B_C4	0.07
B_P3	0.08	B_C3	0.08	B_C4	0.08	A_P2	0.07	B_C3	0.07	B_O2	0.07

For each relevant feature, we counted the total number of times it appeared across the RF classification model for the 2-choice, 4-choice and 11-choice scales. The maximum is 3 for a feature that is relevant for training on all the three fear estimation scales.

The most relevant features for all 3 fear level estimation scales, for the user-independent modality, for Classifier 1 were: B_T8, A_FP1, T_FC6, B_FP1, B_FC5, A_FC6, B_FC6, B_FC2, B_P3, B_C3, HR and GSR. For Classifier C2, the most relevant features were: B_FP1, B_O2, B_FC2, B_C3, B_T8, B_P3, T_FC6, B_FC5, A_FC6, B_FC6, HR, FL_t and GSR. With respect to the user-dependent modality, for Classifier C1, for all 3 fear estimation scales, the most relevant features were: B_P3, B_C3, B_O1, T_FC6, B_FP1, A_FC6, B_P2, A_FP1, B_FC6, B_C4, B_FC2, HR and GSR. With respect to classifier C2, we mention: B_O2, B_FC2, T_FC6, B_FP1, B_FC6, A_FC6, GR, FL_t and GSR.

8. Discussion

The results presented in Tables 4–7 show that the cross-validation and test accuracies obtained after SFS feature selection are lower than those obtained without feature selection. In Table 10, we present the classifiers providing the highest cross-validation and test accuracies for both the player-independent and player-dependent cross-validation and testing methods, on the 2-choice, 4-choice and 11-choice fear scales.

With respect to C1, the classifier predicting fear level, we conclude that the highest cross-validation accuracy (over 98%) was obtained by using either the kNN or RF algorithms, for both the player-independent and player-dependent modalities. The same trend occurs for C2, the classifier predicting the game level to be played next, where very high cross-validation accuracies were recorded by the RF classifier. With respect to the test (or validation) accuracy, for the 2-choice scale, the highest accuracy was obtained by DNN_Model_4 (79.12%) for the player-independent modality and SVM (89.5%) for the player-dependent modality. In the case of the 4-choice scale, the highest accuracies were provided by kNN (52.75%) and SVM (42.5%), respectively. We observed that SVM was very efficient for the player-dependent modality. For the 2-choice scale, both accuracies (79.12% and 89.5%) were higher than the random value of 50% when selecting either 0 or 1 by chance. The same happens in

the case of the 4-choice scale, where the random, "by chance" accuracy is 25%. Both the kNN and SVM classifiers provided an accuracy higher than 25% (kNN—52.75% for the player-independent modality—and SVM—42.5% for the player-dependent method).

Table 10. Highest cross-validation and test accuracies.

Method	**C1**				
	2-Choice Scale		**4-Choice Scale**		**11-Choice Scale**
	Cross-Validation	**Test**	**Cross-Validation**	**Test**	**Cross-Validation**
Player-independent	kNN 99.5% RF 99.25%	DNN_Model_4 79.12%	kNN 99% RF 99%	kNN 52.75%	kNN 98.25% RF 99%
Player-dependent	kNN 99.5% RF 99.75%	SVM 89.5%	kNN 99% RF 99.25%	SVM 42.5%	kNN 98.25% RF 99%
	C2				
	2-Choice Scale		4-Choice Scale		11-Choice Scale
	Cross-Validation	Test	Cross-Validation	Test	Cross-Validation
Player-independent	RF 99.75%	-	RF 100%	-	RF 100%
Player-dependent	RF 99.75%	-	RF 99.75%	-	RF 100%

Both the player-independent and player-dependent training and testing modalities offered good classification results, making it difficult to determine which was best. However, we incline towards using the player-independent one, as we want a more general, less user-specific model.

With respect to features importance (determined by the RF classifier), we observed that GSR, HR and the beta waves play a significant role in fear level prediction for C1. They are followed closely by the alpha and theta activations, but on a lower extent. In the case of C2, the classifier predicting the game level to be played next, the "target fear level" feature, the feature we computed based on the user's current fear level plays a dominant role, not only because it has a high feature importance index determined by the RF classifier, but also because it is selected when using all three fear level estimation scales (2-choice, 4-choice and 11-choice), for both the player-independent and player-dependent modalities. Our findings are in line with the state-of-the-art literature supporting the idea that GSR, HR and the beta waves are related to emotions classification, particularly fear assessment [111,112]. As there are no experiments in which the next game level is predicted based on physiological data, we cannot compare the results obtained by cross-validating and testing C2. However, it is worth pointing out that the same GSR, HR and EEG features are elicited and, in addition, to emphasize the important role that the "target fear level" feature plays in predicting the next level of in-game exposure.

Our results are comparable to those obtained by Liu et al. [12], who used a dynamic difficulty adjustment of game levels based on simple "if" clauses and obtained a classification accuracy of 78%. Having as features both physiological data, Chanel et al. [11] reached a classification accuracy of 63% for the detection of 3 emotional classes in an experiment where 20 participants played a Tetris game with 3 levels of difficulty. Without feature selection, the best classifiers obtained an accuracy of 55% for peripheral features and 48% for EEG features. Feature selection increased the accuracy to 59%, respectively, 56%. Our results are also comparable to those obtained by Lisetti et al. [113], who achieved a classification accuracy of 84% when distinguishing 6 emotional states elicited by movie clips. However, our modality of providing stimuli is more realistic and immersive, as we used for training and testing the classifiers both in vivo and 3D VR stimuli.

9. Conclusions

The purpose of our research was to develop a VR game with ML-based decision support in order to adapt the levels of exposure to the patients' physiological characteristics. To determine the best ML techniques for acrophobia therapy, several classifiers have been trained: Support Vector Machine, Random Forest, k-Nearest Neighbors, Linear Discriminant Analysis and 4 deep neural network models. We proposed two classifiers: one classifier that estimates the current fear level, based on the user's physiological recordings and one that predicts the next exposure scenario, i.e., the game level to be played next. We used 3 scales of measuring fear level, with 2, 4 and 11 possible responses (2-choice, 4-choice and 11-choice scale). The validation accuracy is defined as the measure of similarity between the fear level estimated by the first classifier and the Subjective Unit of Distress reported by the user during gameplay. For the 2-choice scale, the highest accuracy has been obtained by DNN_Model_4 (79.12%) for the player-independent modality and SVM (89.5%) for the player-dependent modality. In the case of the 4-choice scale, the highest accuracies were obtained using kNN (52.75%) and SVM (42.5%), respectively. The cross-validation scores are very high for both classifiers, with the best accuracies obtained by the kNN and RF techniques. The most important features for fear level classification were GSR, HR and the values of the EEG in the beta range. For next game level prediction, the "target fear level", a parameter computed by taking into account the estimated fear level, played a dominant role in classification.

A future study would be to implement a VR-based game for treating other types of phobias. Moreover, we will extend the experiments and involve more subjects, while their physiological responses will be collected and used for training and testing the classifiers. Another direction we will pursue is to perform real-world tests with the 8 acrophobic patients who participated in the current study, expose them to in vivo scenarios and evaluate whether their anxiety levels dropped.

Author Contributions: Conceptualization, O.B. and G.M.; methodology, O.B. and G.M.; software, O.B.; validation, O.B., G.M., A.M., F.M., M.L.; investigation, O.B. and G.M.; resources, A.M. and M.L.; writing—original draft preparation, O.N.; writing—review and editing, O.B. and G.M.; supervision, A.M., M.L. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: The work has been funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125, UEFISCDI project 1/2018 and UPB CRC Research Grant 2017. This work has been funded in part through UEFISCDI, from EEA Grants 2014-2021, project number EEA-RO-NO-2018-0496.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Depression and Other Common Mental Disorders: Global Health Estimates*; World Health Organization: Geneva, Switzerland, 2017.
2. Olesen, J. What Is Fear? Types of Phobias and Their Meanings. 2015. Available online: https://www.fearof.net/what-is-fear-types-of-phobias-and-their-meanings/ (accessed on 20 October 2019).
3. Olesen, J. Phobia Statistics and Surprising Facts about Our Biggest Fears. 2015. Available online: http://www.fearof.net/phobia-statistics-and-surprising-facts-about-our-biggest-fears/ (accessed on 20 October 2019).
4. Uncover the Facts Behind Our Most Common Phobias [Infographic. (2017)]. Available online: https://blog.nationwide.com/common-phobias-statistics/ (accessed on 20 October 2019).
5. Cognitive Behavioral Therapy. 2018. Available online: https://www.psychologytoday.com/us/basics/cognitive-behavioral-therapy (accessed on 20 October 2019).
6. Lamson, R.J. *Virtual Reality Immersion Therapy for Treating Psychological, Psychiatric, Medical, Educational and Self-Help Problems*; U.S. Patent Office: San Rafael, CA, USA, 2002.
7. Fadden, H. Acrophobia (Definition, Causes, Symptoms and Treatment). 2018. Available online: https://www.thehealthyapron.com/acrophobia-definition-causes-symptoms-treatment.html (accessed on 20 October 2019).

8. Opris, D.; Pintea, S.; Garcia-Palacios, A.; Botella, C.; Szamoskozi, S.; David, D. Virtual Reality Exposure Therapy in Anxiety Disorders: A Quantitative Meta-Analysis. *Depress. Anxiety* **2012**, *29*, 85–93. [PubMed]
9. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
10. Calvo, R.; D'Mello, S.; Gratch, J.; Kappas, A. Introduction to Affective Computing. In *The Oxford Handbook of Affective Computing*; Calvo, R., D'Mello, S., Gratch, J., Kappas, A., Eds.; Oxford University Press: Oxford, UK, 2015; pp. 1–8.
11. Chanel, G.; Rebetez, C.; Bétrancourt, M.; Pun, T. Emotion Assessment from Physiological Signals for Adaptation of Game Difficulty. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **2011**, *41*, 1052–1063. [CrossRef]
12. Liu, C.; Agrawal, P.; Sarkar, N.; Chen, S. Dynamic Difficulty Adjustment in Computer Games through Real-Time Anxiety-Based Affective Feedback. *Int. J. Hum.-Comput. Interact.* **2009**, *25*, 506–529. [CrossRef]
13. Bickmore, T.W. Relational Agents in Health Applications: Leveraging Affective Computing to Promote Healing and Wellness. In *The Oxford Handbook of Affective Computing*; Calvo, R., D'Mello, S., Gratch, J., Kappas, A., Eds.; Oxford University Press: Oxford, UK, 2015; pp. 537–558.
14. Messinger, D.S.; Duvivier, L.L.; Warren, Z.E.; Mahoor, M.; Baker, J.; Warlaumont, A.S.; Ruvolo, P. Affective Computing, Emotional Development, and Autism. In *The Oxford Handbook of Affective Computing*; Calvo, R., D'Mello, S., Gratch, J., Kappas, A., Eds.; Oxford University Press: Oxford, UK, 2015; pp. 516–536.
15. North, M.M.; North, S.M.; Coble, J.R. Effectiveness of Virtual Environment Desensitization in the Treatment of Agoraphobia. *Int. J. Virtual Real.* **1995**, *1*, 25–34. [CrossRef]
16. Coelho, C.M.; Waters, A.M.; Hine, T.J.; Wallis, G. The use of virtual reality in acrophobia research and treatment. *J. Anxiety Disord.* **2009**, *23*, 563–574. [CrossRef] [PubMed]
17. Bălan, O.; Moise, G.; Petrescu, L.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Towards a Human-Centered Approach for VRET Systems—Case study for acrophobia. In Proceedings of the 28th International Conference on Information Systems Development, Toulon, France, 28–30 August 2019.
18. Garcia-Palacios, A.; Hoffman, H.G.; See, S.K.; Tsai, A.; Botella, C. Redefining Therapeutic Success with Virtual Reality Exposure Therapy. *CyberPsychol. Behav.* **2001**, *4*, 341–348. [CrossRef]
19. C2Phobia. Available online: https://www.c2.care/en/c2phobia-treating-phobias-in-virtual-reality/ (accessed on 20 October 2019).
20. PSIOUS. Available online: https://www.psious.com/ (accessed on 20 October 2019).
21. Stim Response Virtual Reality. Available online: https://www.biopac.com/application/virtual-reality/ (accessed on 20 October 2019).
22. Virtual Reality Medical Center. Available online: https://vrphobia.com/ (accessed on 20 October 2019).
23. Virtually Better. Available online: www.virtuallybetter.com (accessed on 20 October 2019).
24. VR Treatment Program at Duke University School of Medicine. Available online: https://psychiatry.duke.edu/virtual-reality-therapy-phobias (accessed on 20 October 2019).
25. Bravemind. Available online: http://medvr.ict.usc.edu/projects/bravemind/ (accessed on 20 October 2019).
26. Limbix. Available online: https://www.limbix.com/ (accessed on 20 October 2019).
27. Phobos. Available online: https://samsungvr.com/view/Uu9ME9YXR_B (accessed on 20 October 2019).
28. Activity. Available online: https://www.unitylab.de/ (accessed on 20 October 2019).
29. Oculus Rift. Available online: https://www.oculus.com/?locale=en_US (accessed on 20 October 2019).
30. Microsoft Kinect. Available online: https://developer.microsoft.com/en-us/windows/kinect (accessed on 20 October 2019).
31. Coelho, C.M.; Silva, C.F.; Santos, J.A.; Tichon, J.; Wallis, G. Contrasting the Effectiveness and Efficiency of Virtual Reality and Real Environments in the Treatment of Acrophobia. *Psychol. J.* **2008**, *6*, 203–216.
32. Emmelkamp, P.M.; Krijn, M.; Hulsbosch, A.M.; de Vries, S.; Schuemie, M.J.; van der Mast, C.A. Virtual Reality Treatment Versus Exposure in Vivo: A comparative Evaluation in Acrophobia. *Behav. Res. Ther.* **2002**, *40*, 509–516. [CrossRef]
33. HTC Vive. Available online: https://www.vive.com/eu/ (accessed on 20 October 2019).
34. The Climb. Available online: http://www.theclimbgame.com/ (accessed on 20 October 2019).
35. Ritchie's Plank Experience. Available online: https://www.viveport.com/apps/9347a360-c6ea-4e35-aaf1-9fab4f41cb79/Richie\T1\textquoterights_Plank_Experience/ (accessed on 20 October 2019).
36. Arachnophobia. Available online: https://store.steampowered.com/app/485270/Arachnophobia/ (accessed on 20 October 2019).

37. Limelight. Available online: https://store.steampowered.com/app/544880/Limelight_VR/ (accessed on 20 October 2019).
38. Samsung Fearless Cityscapes. Available online: https://www.oculus.com/experiences/gear-vr/821606624632569/ (accessed on 20 October 2019).
39. Samsung Fearless Landscapes. Available online: https://www.oculus.com/experiences/gear-vr/1290835750988761/ (accessed on 20 October 2019).
40. Samsung Gear V.R. Available online: https://www.samsung.com/global/galaxy/gear-vr/ (accessed on 20 October 2019).
41. Samsung Gear S2. Available online: https://www.samsung.com/global/galaxy/gear-s2/ (accessed on 20 October 2019).
42. Emotion Definition. Available online: https://en.oxforddictionaries.com/definition/emotion (accessed on 20 October 2019).
43. Koelstra, S.; Muehl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
44. Solomon, C.R. Emotion. Available online: https://www.britannica.com/science/emotion#ref283140 (accessed on 20 October 2019).
45. Ekman, P.; Sorenson, E.R.; Friesen, W.V. Pan-Cultural Elements in Facial Displays of Emotion. *Science* **1969**, *164*, 86–88. [CrossRef]
46. Ekman, P.; Cordaro, D. What is Meant by Calling Emotions Basic. *Emot. Rev.* **2011**, *3*, 364–370. [CrossRef]
47. Plutchik, R. Emotion: Theory, Research, and Experience. In *Theories of Emotion: 1*; New York Academic: New York, NY, USA, 1980; Volume 1.
48. Sacharin, V.; Schlegel, K.; Scherer, K.R. *Geneva Emotion Wheel Rating Study (Report)*; University of Geneva, Swiss Center for Affective Sciences: Geneva, Switzerland, 2012.
49. Russell, J.A. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
50. Bontchev, B. Adaptation in Affective Video Games: A Literature Review. *Cybern. Inf. Technol.* **2016**, *16*, 3–34. [CrossRef]
51. Steimer, T. The Biology of Fear and Anxiety-Related Behaviors. *Dialogues Clin. Neurosci.* **2002**, *4*, 231–249.
52. Golumbic-Zion, E. What Is EEG? Available online: https://www.mada.org.il/brain/articles/faces-e.pdf (accessed on 20 October 2019).
53. Solomon, C.R. The Neurobiology of Emotion. Available online: https://www.britannica.com/science/emotion#ref283143 (accessed on 20 October 2019).
54. Davidson, R.J. Cerebral Asymmetry and Emotion: Conceptual and Methodological Conundrums. *Cognit. Emot.* **1993**, *7*, 115–138. [CrossRef]
55. Davidson, R.J. Affective Style and Affective Disorders: Perspectives from Affective Neuroscience. *Cognit. Emot.* **1998**, *12*, 307–330. [CrossRef]
56. Coan, J.A.; Allen, J.J.B.; Harmon-Jones, E. Voluntary Facial Expression and Hemispheric Asymmetry over the Frontal Cortex. *Psychophysiology* **2001**, *38*, 912–925.
57. Coan, J.A.; Allen, J.J.B. Frontal EEG Asymmetry and the Behavioral Activation and Inhibition Systems. *Psychophysiology* **2003**, *40*, 106. [CrossRef]
58. Park, K.S.; Choi, H.; Lee, K.J.; Lee, J.Y.; An, K.O.; Kim, E.J. Emotion Recognition Based on the Asymmetric Left and Right Activation. *Int. J. Med. Med Sci.* **2011**, *3*, 201–209.
59. Wise, V.; McFarlane, A.C.; Clark, C.R.; Battersby, M. An Integrative Assessment of Brain and Body Function 'at rest' in Panic Disorder: A Combined Quantitative EEG/Autonomic Function Study. *Int. J. Psychophysiol.* **2011**, *79*, 155–165. [CrossRef]
60. Knott, V.; Lapierre, Y.D.; Fraser, G.; Johnson, N. Auditory Evoked Potentials in Panic Disorder. *J. Psychiatry Neurosci.* **1991**, *16*, 215–220. [PubMed]
61. Engelbregt, H.J.; Keeser, D.; Promes, V.H.; Verhagen-Schouten, S.; Deijen, J.B. In-Vivo EEG Changes During a Panic Attack in a Patient with Specific Phobia. *J. Med. Cases* **2012**, *3*, 34–38. [CrossRef]
62. Gordeev, S.A. Clinical and Psychophysiological Study of Patients with Panic Attacks with or without Agoraphobic Disorders. *Zh. Nevrol. Psikhiatr. Im. S. S. Korsakova.* **2007**, *107*, 54–58. [CrossRef] [PubMed]
63. Knyazev, G.G.; Savostyanov, A.N.; Levin, E.A. Alpha Oscillations as a Correlate of Trait Anxiety. *Int. J. Psychophysiol.* **2004**, *53*, 147–160. [CrossRef]

64. Suetsugi, M.; Mizuki, Y.; Ushijima, I.; Kobayashi, T.; Tsuchiya, K.; Aoki, T.; Watanabe, Y. Appearance of Frontal Midline Theta Activity in Patients with Generalized Anxiety Disorder. *Neuropsychobiology* **2000**, *41*, 108–112. [CrossRef]
65. Schutter, D.J.; Van Honk, J. A eElectrophysiological Ratio Markers for the Balance between Reward and Punishment. *Cogntive Brain Res.* **2005**, *24*, 685–690. [CrossRef]
66. Putman, P.; Van, P.J.; Maimari, I.; Vander, W.S. EEG Theta/Beta Ratio in Relation to Fear-Modulated Response-Inhibition, Attentional Control, and Affective Traits. *Biol. Psychol.* **2010**, *83*, 73–78. [CrossRef]
67. Choi, J.S.; Bang, J.W.; Heo, H.; Park, K.R. Evaluation of Fear Using Nonintrusive Measurement of Multimodal Sensors. *Sensors* **2015**, *15*, 17507–17533. [CrossRef]
68. Cheemalapati, S.; Adithya, P.C.; Valle, M.D.; Gubanov, M.; Pyayt, A. Real Time Fear Detection Using Wearable Single Channel Electroencephalogram. *Sensor Netw. Data Commun.* **2016**, *5*, 140. [CrossRef]
69. Petrantonakis, P.G.; Hadjileontiadis, L.J. EEG-based Emotion Recognition Using Hybrid Filtering and Higher Order Crossings. In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–6.
70. Carvalho, M.R.; Velasques, B.B.; Cagy, M.; Marques, J.B.; Teixeira, S.; Nardi, A.E.; Piedade, R.; Ribeiro, P. Electroencephalographic Findings in Panic Disorder. *Trends Psychiatry Psychother.* **2013**, *35*, 238–251. [CrossRef]
71. Rainville, P.; Bechara, A.; Naqvi, N.; Damasio, A.R. Basic Emotions are Associated with Distinct Patterns of Cardiorespiratory Activity. *Int. J. Psychophysiol.* **2006**, *61*, 5–18. [CrossRef]
72. Gouizi, K.; Reguig, F.B.; Maaoui, C. Analysis Physiological Signals for Emotion Recognition. In Proceedings of the 2011 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), Tipaza, Algeria, 9–11 May 2011; pp. 147–150.
73. Muhl, C.; Heylen, D. Cross-Modal Elicitation of Affective Experience. In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2019; pp. 1–2.
74. Agrafioti, F.; Hatzinakos, D.; Anderson, A. ECG Pattern Analysis for Emotion Detection. *IEEE Trans. Affect. Comput.* **2012**, *3*, 102–115. [CrossRef]
75. Healey, J.A. Affect Detection in the Real World: Recording and Processing Physiological Signals. In Proceedings of the IEEE 3rd International Conference on Affective Computing and Intelligent Interaction, Amsterdam, The Netherlands, 10–12 September 2009; IEEE Press: Amsterdam, The Netherlands, 2009; Volume 1, pp. 729–734.
76. Fleureau, J.; Guillotel, P.; Huynh-Thu, Q. Physiological-based Affect Event Detector for Entertainment Video Applications. *IEEE Trans. Affect. Comput.* **2012**, *3*, 379–385. [CrossRef]
77. AlZoubi, O.; D'Mello, S.K.; Calvo, R.A. Detecting Naturalistic Expressions of Nonbasic Affect using Physiological Signals. *IEEE Trans. Affect. Comput.* **2010**, *3*, 298–310. [CrossRef]
78. Westerink, J.; Ouwerkerk, M.; de Vries, G.-J.; de Waele, S.; van den Eerenbeemd, J.; van Boven, M. Emotion Measurement Platform for Daily Life Situations. In Proceedings of the International Conference (ACII 2009), Amsterdam, The Netherlands, 10–12 September 2009.
79. Wiederhold, B.K.; Jang, D.P.; KIM, S.I.; Wiederhold, M.D. Physiological Monitoring as an Objective Tool in Virtual Reality Therapy. *Cyberpsychol. Behav.* **2002**, *5*, 77–82. [CrossRef]
80. Peterson, S.M.; Furuichi, E.; Ferris, D.P. Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading. *PLoS ONE* **2018**, *13*, 1–17. [CrossRef]
81. Kritikos, J.; Tzannetos, G.; Zoitaki, C.; Poulopoulou, S.; Koutsouris, D. Anxiety detection from Electrodermal Activity Sensor with movement & interaction during Virtual Reality Simulation. In Proceedings of the 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA, 20–23 March 2019.
82. Lee, M.; Bruder, G.; Welch, G. The Virtual Pole: Exploring Human Responses to Fear of Heights in Immersive Virtual Environments. *J. Virtual Real. Broadcast.* **2017**, *6*, 1–14.
83. Wang, X.W.; Nie, D.; Lu, B.L. Emotional State Classification from EEG Data Using Machine Learning Approach. *Neurocomputing* **2014**, *129*, 94–106. [CrossRef]
84. Picard, R.W. Affective Computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321. 1995. Available online: http://vismod.media.mit.edu/tech-reports/TR-321.pdf (accessed on 20 October 2019).

85. Nasoz, F.; Alvarez, K.; Lisetti, C.L.; Finkelstein, N. Emotion Recognition from Physiological Signals Using Wireless Sensors for Presence Technologies. *Cognit. Technol. Work* **2004**, *6*, 4–14. [CrossRef]
86. Jirayucharoensak, S.; Pan-Ngum, S.; Israsena, P. EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. Hindawi Publishing Corporation. *Sci. World J.* **2014**, *2014*, 627892. [CrossRef]
87. Fazi, M.B. Can a Machine Think (Anything New)? Automation Beyond Simulation. *AI Soc.* **2019**, *34*, 813–824. [CrossRef]
88. Jerritta, S.; Murugappan, M.; Nagarajan, R.; Wan, K. Physiological Signals Based Human Emotion Recognition: A Review. In Proceedings of the IEEE 7th International Colloquium on Signal Processing and its Applications, Penang, Malaysia, 4–6 March 2011. [CrossRef]
89. Jonghwa, K.; Ande, E. Emotion Recognition Based on Physiological Changes in Music Listening. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 2067–2083.
90. Wan-Hui, W.; Yu-Hui, Q.; Guang-Yuan, L. Electrocardiography Recording, Feature Extraction and Classification for Emotion Recognition. In Proceedings of the WRI World Congress on Computer Science and Information Engineering Los Angeles, Los Angeles, CA, USA, 31 March–2 April 2009.
91. Caridakis, G.; Castellano, G.; Kessous, L.; Raouzaiou, A.; Malatesta, L.; Asteriadis, S.; Karpouzis, K. Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech. In *Artificial Intelligence and Innovations 2007: From Theory to Applications*; Boukis, C., Pnevmatikakis, L., Polymenakos, L., Eds.; Springer: Boston, MA, USA, 2007; Volume 247, pp. 375–388.
92. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15), New York, NY, USA, 9–13 November 2015; pp. 443–449. [CrossRef]
93. Teo, J.; Chew, L.H.; Chia, J.T.; Mountstephens, J. Classification of Affective States via EEG and Deep Learning. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*. [CrossRef]
94. Liu, C.; Rani, P.; Sarkar, N. An Empirical Study of Machine Learning Techniques for Affect Recognition in Human-Robot Interaction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 1–8. [CrossRef]
95. Bălan, O.; Moise, G.; Petrescu, L.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Emotion Classification Based on Biophysical Signals and Machine Learning Techniques. *Symmetry* **2020**, *12*, 21. [CrossRef]
96. Ray, L.C.; Fukuoka, Y. Machine Learning and Therapeutic Strategies in VR. In Proceedings of the 9th International Conference on Digital and Interactive Arts (Artech 2019), Braga, Portugal, 23–25 October 2019.
97. Hu, F.; Wang, H.; Chen, J.; Gong, J. Research on the characteristics of acrophobia in virtual altitude environment. In Proceedings of the 2018 IEEE International Conference on Intelligence and Safety for Robotics, Shenyang, China, 24–27 August 2018.
98. Šalkevicius, J.; Damaševičius, R.; Maskeliunas, R.; Laukienė, I. Anxiety Level Recognition for Virtual Reality Therapy System Using Physiological Signals. *Electronics* **2019**, *8*, 1039. [CrossRef]
99. Bălan, O.; Moise, G.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Automatic Adaptation of Exposure Intensity in VR Acrophobia Therapy, Based on Deep Neural Networks. In Proceedings of the Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden, 8–14 June 2019.
100. Bălan, O.; Moise, G.; Moldoveanu, A.; Moldoveanu, F.; Leordeanu, M. Does Automatic Game Difficulty Level Adjustment Improve Acrophobia Therapy? Differences from Baseline. In Proceedings of the VRST '18, Tokyo, Japan, 28 November–1 December 2018. [CrossRef]
101. Huppert, D.; Grill, E.; Brandt, T. A New Questionnaire for Estimating the Severity of Visual Height Intolerance and Acrophobia by a Metric Interval Scale. *Front. Neurol.* **2017**, *8*, 211. [CrossRef]
102. Acticap Xpress Bundle. Available online: https://www.brainproducts.com/productdetails.php?id=66 (accessed on 20 October 2019).
103. Shimmers Multisensory. Available online: http://www.shimmersensing.com/ (accessed on 20 October 2019).
104. Shimmer3 GSR+ Unit. Available online: https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor (accessed on 20 October 2019).
105. Scikit Learn Python Library. Available online: http://scikit-learn.org (accessed on 20 November 2018).
106. Tensor Flow Python Framework. Available online: https://www.tensorflow.org/ (accessed on 20 October 2019).
107. Keras Library. Available online: https://keras.io/ (accessed on 20 October 2019).

108. Unity Game Engine. Available online: https://unity.com/ (accessed on 20 October 2019).
109. Open Vibe. Available online: http://openvibe.inria.fr/ (accessed on 20 October 2019).
110. Lab Stream Layer. Available online: https://github.com/sccn/labstreaminglayer (accessed on 20 October 2019).
111. Arikan, K.; Boutros, N.N.; Bozhuyuk, E.; Poyraz, B.C.; Savrun, B.M.; Bayar, R.; Gunduz, A.; Karaali-Savrun, F.; Yaman, M. EEG Correlates of Startle Reflex with Reactivity to Eye Opening in Psychiatric Disorders: Preliminary Results. *Clin. EEG Neurosci.* **2006**, *37*, 230–234. [CrossRef]
112. Kometer, H.; Luedtke, S.; Stanuch, K.; Walczuk, S.; Wettstein, J. The Effects Virtual Reality Has on Physiological Responses as Compared to Two-Dimensional Video. *J. Adv. Stud. Sci.* **2010**, *1*, 1–21.
113. Lisetti, C.L.; Nasoz, F. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *J. Appl. Signal Process.* **2004**, *11*, 1672–1687. [CrossRef]

Towards a Human-Centered Approach for VRET Systems: case study for acrophobia

Oana Bălan
University POLITEHNICA of Bucharest
Bucharest, Romania *oana.balan@cs.pub.ro*

Ștefania Cristea
University POLITEHNICA of Bucharest
Bucharest, Romania *stefania.cristea1708@gmail.com*

Alin Moldoveanu
University POLITEHNICA of Bucharest
Bucharest, Romania *alin.moldoveanu@cs.pub.ro*

Gabriela Moise
Petroleum-Gas University of Ploiești
Ploiești, Romania *gmoise@upg-ploiesti.ro*

Marius Leordeanu
University POLITEHNICA of Bucharest
Bucharest, Romania *marius.leordeanu@cs.pub.ro*

Florica Moldoveanu
University POLITEHNICA of Bucharest
Bucharest, Romania *florica.moldoveanu@cs.pub.ro*

Abstract

This paper presents a human-centered methodology for designing and developing Virtual Reality Exposure Therapy (VRET) systems. By following the steps proposed by the methodology – Users analysis, Domain Analysis, Task Analysis and Representational Analysis, we developed a system for acrophobia therapy composed of 9 functional, interrelated modules which are responsible for patients, scenes, audio and graphics management, as well as with physiological monitoring and event triggering. The therapist visualizes in real time the patient's biophysical signals and adapts the exposure scenario accordingly, as. he can lower or increase the level of exposure. There are 3 scenes in the game, depicting a ride by cable car, one by ski lift and a walk by foot in a mountain landscape. A reward system is implemented and emotion dimension ratings are collected at predefined points in the scenario. They will be stored and later used for constructing an automatic machine learning emotion recognition and exposure adaptation module.
Keywords: Virtual Reality, Exposure Therapy, Human-Centered, Acrophobia, Phobia

1. Introduction

Phobia is a prevalent anxiety disorder of our times, affecting 13% of the world's population. They are characterized by an extreme fear of objects or situations, distressing panic attacks and physical symptoms such as sweating, trembling, rapid heartbeat, headaches, dizziness, confusion and disorientation. In severe situations, some people experience psychological symptoms such as fear of losing control or even fear of dying. Phobias are divided into 3 categories – social phobias (fear of meeting people of higher authority, using a telephone or speaking before a large crowd), agoraphobia (fear of open spaces) and specific phobias, which are generated by specific objects or situations. In what concerns social phobias, they affect

people of all ages, but usually appear in adolescence. 45% of people with social phobias develop agoraphobia and the fear of having an anxiety attack in public or embarrassing themselves, while 17% develop depression [25]. 15-20% of the world's population experience specific phobias at least once in the lifetime [23]. At world level, specific phobias have the following prevalence: acrophobia (fear of height) – 7.5%, arachnophobia (fear of spiders) – 3.5%, aerophobia (fear of flying) – 2.6%, astraphobia (fear of lightning and thunder) – 2.1%, dentophobia (fear of dentist) – 2.1% [22]. The annual total costs of social phobia were 11.952 euros in the Netherlands, higher than the total costs for people with no mental disorder, of 2957 euros [1]. As concerns the European Union, the direct (diagnosis and treatment) and indirect (invisible costs associated with income losses due to mortality and disability) costs were estimated at 798 billion euros. They are expected to double by 2030 [40].

Of the people suffering from social phobias, only 23% seek specialized help. 80% of the patients turn to medicines and Cognitive Behavior Therapy (CBT), a method of gradual in-vivo exposure to stimuli and thought control. Unfortunately, only 50% of the persons suffering from social phobias and 20% of those affected by specific phobias recover completely [25].

Besides CBT and in-vivo exposure, a new therapy has emerged, namely *VRET* (*Virtual Reality Exposure Therapy*). The user is presented a computer-generated virtual environment, either on a desktop or mobile platform, via a Head Mounted Display (HMD). Virtual environments can be easily controlled by the therapist, customized and adapted to the condition of each patient. They are immersive, appealing, cheap and most importantly, safe. Over 80% of the patients prefer virtual exposure therapy over the classical in-vivo exposure [11]. VRET has a strong stability of results in time, equal to that obtained by CBT therapy [24]. However, it is appropriate for people who do not possess high imaginative skills such as those required for CBT. It also provides a more comfortable sensation than in-vivo exposure, knowing that it is only a virtual immersion from which you can abscond as soon as you feel like losing control.

In this paper we propose a methodology inspired from the HCDID model proposed by [45] and from the NADI model of Mieke and Dorst [42] for designing and developing a VRET system. In addition, we provide a case study for acrophobia therapy. Such, we designed a virtual environment illustrating a mountain scenario where the user can ride by cable car, ski lift and walk by foot. The therapist can manage the patients, visualize their physiological parameters and adapt the scenario accordingly. The design is human-centered, thus it meets both the patients' and therapists' requirements. In this phase of research, we collect data from the users and from the therapists (biophysical signals, actions performed in the virtual environment, user behavior, general performance, the modality in which the clinical specialist reacts to the patient's performance and physiological data, adapting the exposure scenario accordingly). This data will be used for constructing a computational model with various feature extraction and machine learning techniques that will automatically recognize human emotions and adapt the virtual exposure in real time.

The paper is organized as follows: chapter 2 presents existing systems for phobia therapy, chapter 3 describes the emotion models, chapter 4 presents the human-centered paradigm, chapter 5 is dedicated to our proposed human-centered VRET system design methodology and chapter 6 introduces a case study, the development of a VRET system for acrophobia treatment. Finally, we show the study's conclusions and provide future directions of research.

2. Virtual Reality systems for phobia therapy

In order to perform a comprehensive analysis of the existing Virtual Reality (VR) systems for phobia therapy, we considered 3 main categories: *platforms*, *applications for desktop and mobile devices* and *systems developed within an academic research*.

2.1. Platforms

C2Phobia software [4] is composed of more than 70 configurable exposure environments (the therapist can add/remove elements from the environment) for treating a wide range of phobic conditions: Acrophobia, Agoraphobia, Claustrophobia, Ochlophobia, Arachnophobe, Aviophobia, School phobia, Fear of public speaking, Fear of pigeons, Fear of dogs, Fear of cats, Fear of the hospital. The patient is exposed gradually to different levels of anxiety according to his pathologies. PSIOUS [29] provides over 50 resources (VR and augmented reality environments, 360º videos) with real-time view of what the patient is seeing during the session. Stim Response Virtual Reality [38] offers fully modular environments for acrophobia, fear of flying and fear of public speaking therapy. The events from the virtual or augmented world and the physiological data (ECG, EEG, EOG, EMG, EGG, EDA, temperature,

respiration, pulse) are synchronized. Virtual Reality Medical Center (VRMC) [43] uses VRET in combination with biofeedback and CBT to treat phobias (fear of flying, fear of driving, fear of heights, fear of public speaking, claustrophobia, agoraphobia), anxiety (including pre-surgical anxiety), stress and chronic pain. This system is used also for treating post-traumatic stress disorder caused by military deployment. Each stage can be repeated until the client feels comfortable. At every step, the therapist can see and hear what the client is experiencing. If the level of anxiety becomes too high, the user can return to a lower level or exit the virtual world. Virtually Better [44] offers Bravemind, a system for alleviating the psychological repercussions of war for the soldiers who served in Iraq or Afghanistan. Bravemind is accompanied by vibrotactile feedback (sensations associated with engine rumbling, explosions, firefights), ambient noises and scent machines. Limbix [18] offers VR environments built from panoramic images and videos. The scenes are interactive, so that the therapists can change them in real-time. PHOBOS [26] is designed for individuals, professionals and organizations. It ensures gradual exposure to stimuli and interactive 3D environments that address agoraphobia, social anxiety disorders and specific phobias.

2.2. Applications for desktop and mobile devices

For acrophobia therapy, the most popular games are The Climb [39], Ritchie's Plank Experience [31], Samsung Fearless Cityscapes [33] for Gear VR, Samsung Fearless Landscapes [34]. In Arachnophobia [3], the user looks at specific spots on a piece of paper in front of him and is able to control the amount of exposure to virtual spiders. Limelight [19] for HTC Vive puts the user on stage in front of a virtual crowd that can change its mood and behavior. For treating fear of public speaking, he can give presentations in business meetings, small classrooms or large halls.

2.3. Systems developed within the academic context

Acrophobia Therapy with Virtual Reality (AcTiVity-System) [2] uses Oculus Rift to render the 3D scenes, a Microsoft Kinect for motion tracking and a heart rate sensor. A large experiment, involving 100 users, took place in order to evaluate the system and the results showed that all the participants in the VR group recorded a certain level of fear reduction, with the average reduction being 68%. Half of the participants in the VR group had a reduction in fear of heights by over three quarters. VR Phobias [5] contains a static environment depicting the view of a hotel balcony. The results of an experiment involving 15 users showed the same rates of success for the users treated in a virtual environment and for those exposed to a real-world environment. However, the virtual sessions were shorter (22 minutes), compared to the real-world ones (51 minutes). The acrophobia system developed at University of Amsterdam and Delft University of Technology [9] comprises three different virtual environments: a mall, a fire escape and a roof garden. 29 patients have been exposed to these virtual environments in the presence of the therapist. At the end of the experiment, the subjects have reduced their anxiety and avoidance levels.

3. Emotion models

Some of the most challenging subjects in psychology are related to emotions, emotional-eliciting stimuli and the modalities of measuring affective changes. There are many theories of emotion, with each author offering his own perspective on the topic. In 1969, Izard concluded that *the area of emotional experience and behaviour is one of the most confused and ill-defined in psychology* [14].

Emotions have a complex and multi-aspect nature. According to H. Hockenbury & E. Hockenbury, emotion is seen *as a complex psychological state that involves three distinct components: a subjective experience, a physiological response and a behavioral or expressive response* [13]. While a review on emotion literature in psychology is beyond the scope of this paper, we adopt the definition proposed by H. Hockenbury & E. Hockenbury and present the most relevant emotion models and key concepts used in emotion recognition.

Regarding the emotion models, there are mainly two perspectives: *discrete and dimensional*. In the discrete model, it is assumed the existence of a basic set of emotions. Ekman and Friesen identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [7]. Later, the list was updated including embarrassment, excitement, contempt, shame, pride, satisfaction, amusement, guilt, relief, wonder, ecstasy and sensory pleasure [8]. In the dimensional model, an emotion is described by two or three dimensions, which represent fundamental properties. Russel suggested in his circumplex model of affect the usage of two dimensions: *the arousal or activation dimension* to express the intensity of emotion and *the*

valence dimension to express the way in which the emotion is felt, either positive or negative [32]. *Dominance* was related to the extent to which a person can control his behavior. Nowadays, valence, arousal and dominance are still used as three basic dimensions to express the emotional states. Each discrete emotion can be viewed as a combination of two or three dimensions [28], [21]. For example, *fear is characterized by negative valence, high arousal and low dominance*.

Many laboratory experiments have been carried out in order to study emotions. In [41], a comparative study regarding the capacities of pictures and films to induce emotions is provided. The Self-Assessment Manikin scale was used to rate the emotion and arousal states [17]. The results obtained were unexpected: films were less effective than pictures stimuli. Two stimuli were used in [20] to induce emotional states: self-induced emotional imagery and audio/video clips. Electroencephalography (EEG) brain signals were automatically analysed and used to recognise human emotions. Facial expressions, posture, voice, body motion reflect emotional states [8], [7], [21], [27], [30]. With the development of technology, various data could be acquired and processed, thus leading to automatic emotion recognition systems development. The best performance is achieved by multi-modal emotion recognition.

4. The Human-Centered Paradigm

Nowadays, we are witnessing the explosion of the *Human-Centered paradigm*. There are many definitions which attempt to encompass various aspects of *human-centered.* We find *human-centered* related to with different concepts such as computing, design, systems, machine learning, software engineering and so forth.

In the final report of the workshop Human-Centered Systems (HCS): Information, Interactivity, and Intelligence, 1997, the participants agreed and defined the human-centered systems as *an emerging discipline that combines principles from cognitive science, social science, computer science and engineering to study the design space of systems that support and expand fundamental human activities* [10]. Jaimes et al. notice that the aim of Human-Centered Computing (HCC) is the tight integration of human sciences and computer science to build computing systems with a human focus from beginning to end [15].

Human-centered Machine Learning (HML) proposes a new approach for Machine Learning (ML) algorithms. They consider human goals and contexts in designing ML algorithms, so that ML becomes more useful and usable [12]. The human and the computer have to adapt to each other: *the human can change the behavior of the machine and the machine can change the human's goals*. Applied ML is seen as a co-adaptive process with the computers being part of human design process [12].

Generally speaking, the *Human-Centered Design* (*HCD*) deals with those methods and principles used to design and develop any types of services or products for people, taking into account the utility, pleasure and meaning parameters [42]. Mieke and Dorst developed the NADI model based on four layers of human Needs and Aspirations for Application in a Design and Innovation process [42]:

I Solutions – shows what the people want or need

II Scenarios – describes how the people interact with a solution in a specific context of use

III Goals and IV Themes – both describe why the people want or need certain solutions.

The goals take into account the context of the problem, while themes deal with the context-free analysis of underlying needs and aspirations.

In 2003, Seffah and Andreevskaia noted that the HCD techniques are still insufficiently integrated in software engineering methodologies [36]. Considering the movement of software engineering from the traditional software development to the human-centered development, they proposed the following process features: *user-driven; solution focus; multidisciplinary teamwork including users, customers, human factor experts; focus on external attribute; quality defined be user satisfaction and performance; implementation of user-validated solution only; understanding the context of use* [25].

In [45] a Human-Centered Distributed Information Design (HCDID) methodology is introduced. HCDID comprises two related parts: the first part includes multiple levels of analysis for single user human-centered design (*user, functional, representational, and task analysis*); the second part is dedicated to additional analysis for designing distributed human-centered systems.

5. A Human-Centered VRET System Design Methodology

VRET systems comprise various technologies: VR, AR and ML. Related to VR, Jerald noted in his book that *We must create VR experiences with both emotion and logic* [16]. In our methodology for Human-Centered VRET (HCVRET) systems development (Fig. 1), we use a layers-based analysis adopted from the HCDID model proposed by [45] and from the NADI model of Mieke and Dorst [42]. For the HCVRET implementation, we consider the dimensional model of emotions.

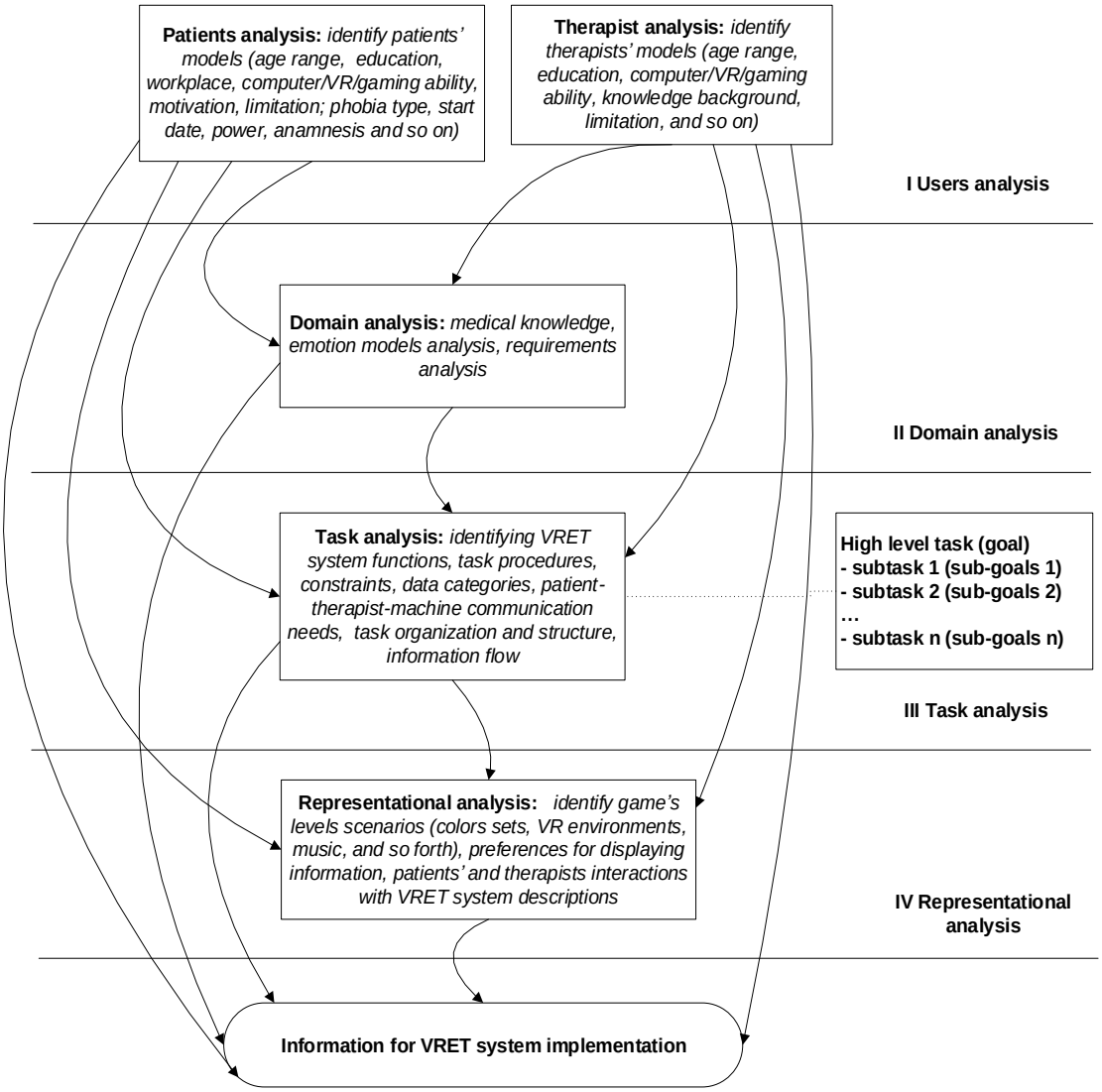

Fig. 1. Layers-based analysis for designing a human-centered VRET systems (adapted after [45])

The layers-based analysis for designing a human-centered VRET system comprises 4 levels. *Level 1* is dedicated to the users' analysis: patients and therapists. Users' patterns and features of these patterns are identified at this stage. It is important to know their medical history, motivation, education, data about the phobia condition. We are interested in the gaming and computer abilities of the patients, as our intention is to develop a game-based VRET system. The therapists also use the system. They supervise the therapy and can intervene during the game. *Level 1* provides information to the following levels. *Level 2* deals with the system analysis requirements, emotion models and knowledge about the mental illness. All information is modeled and encoded to be computationally processed. The VRET system contains a series of tasks which are undertaken by the patients in the therapy. All tasks and subtasks are analyzed at *Level 3*. Each task has a hierarchical structure: high level tasks related to a goal and subtasks related to sub-goals. Also, there are defined the tasks performed by the therapists: for example, the task of setting the next game level in the therapy. The patients, the therapists and the machines need to communicate in a simple and efficient way. Task analysis involves defining the work procedures. An example of procedure is: the patient plays no more than 15 minutes followed by a relaxation period of 10 minutes. In this way, the game-based VRET system is designed to be adaptable to the model of the patient. At *Level 4*, we identify the patients' and therapists' preferences for colors or sounds, for a certain game, for urban or natural landscapes, for certain technologies and so on. All the information acquired at this stage

is used for implementing the VRET.

6. Case study: development of a VRET system for acrophobia treatment

6.1. Software architecture

In this chapter we present the development of a VRET system for treating acrophobia. Our virtual environment is rendered via HTC Vive and depicts a natural setting (a mountain scene with hills and valleys, peaks, forests, a lake, river, transparent platform above a canyon and a transparent bridge) during daytime. The VR environment has been developed using the C# programming language and the Unity graphics engine. The software architecture is composed of the following modules:

Users Manager (Fig. 2) – manages the patients, being dedicated to the therapist. New patients can be added to the system and information about them introduced – name, age, height, sex. Also, each patient selects at this stage his favorite song / picture /quote, which will be presented during the virtual exposure whenever he considers that he needs to relax and calm down. This module also manages existing users, replays sessions and allows the therapist to see statistics concerning the patients' performance. The Users Manager module is connected to a SQL database that stores all the participants' data.

Fig. 2. Users Manager interface

Resources Manager - loads and caches all resources needed at runtime (the patient's profile, scenes, game objects, assets, etc.)

Graphics Manager – ensures graphics rendering and processing, input & output windows and UI (User Interface) display

Input Manager – manages user input from the HTC Vive controllers. The patient interacts with the virtual environment – displacement, objects selection, menu selection, buttons pressing via the HTC Vive controllers.

Audio Manager – audio rendering management: environmental soundscapes (birds chirping, the sound of the wind), auditory icons when the user selects something from the menu, enters or exits the game, audio cues, plays the user's favorite music clip whenever he needs to relax and take a break from the virtual exposure

Scenes Manager – manages the scenes (Fig. 3). The user can select any of the following 3 scenes: a ride by cable car (Fig. 4, Fig. 5), a ride by ski lift (Fig. 6) and a walk by foot (Fig. 7, Fig. 8). Throughout any of these routes, there are 10 stop points where a mathematical quiz is applied in order to detach the patient from the virtual exposure, deactivate the right brain hemisphere responsible with emotional processing and activate the left one which manages logical and rational responses. After the user correctly answers the mathematical question (Fig. 9), he is required to select his valence (Fig. 10), arousal (Fig. 11) and dominance (Fig. 12) levels using Self-Assessment Manikins. If he does not correctly answer the current mathematical question, another one appears on the screen and the process is repeated. At the end of the route, the user is returned to the main menu to select another ride, if he wants. At each moment of time, he can stop the cable car or ski lift from moving, as well as to get down and return to the main menu. At any time, the user can choose to take a pause to relax and listen to his favorite piece of music, see a photo depicting something he enjoys and read his favorite quote.

Fig. 3. Start menu

Fig. 4. View from the cable car

Fig. 5. The cable car

Fig. 6. View from the ski lift

Fig. 7. View from the glass platform

Fig. 8. The glass platform

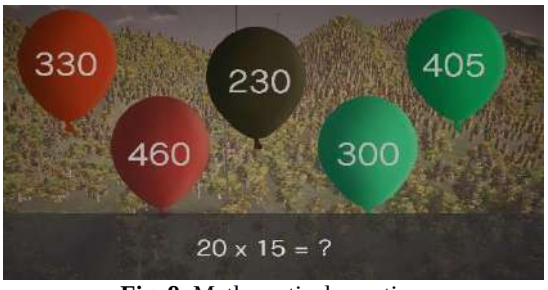

Fig. 9. Mathematical question

Fig. 10. Valence rating

Fig. 11. Arousal rating

Fig. 12. Dominance rating

Physiological Monitoring Module – records physiological data (heart rate (HR) and galvanic skin response GSR)) (Fig. 13). High HR and increased GSR (skin conductivity) are associated with anxiety and fear. Both the user and the therapist can visualize and monitor these parameters and the therapist can also modify the patient's exposure level whenever he considers that the biophysical signals exceeded a critical threshold (Fig. 14).

Fig. 13. GSR and HR recording

Fig. 14. User playing with HTC Vive

Event System Module - triggers various actions during gameplay like saving statistics, recording valence/arousal/dominance rates, rendering events, animation events and any kind of communication between completely various modules or game objects.

Game Manager – integrates and operates all the modules mentioned above.

The software architecture is presented in Fig. 15.

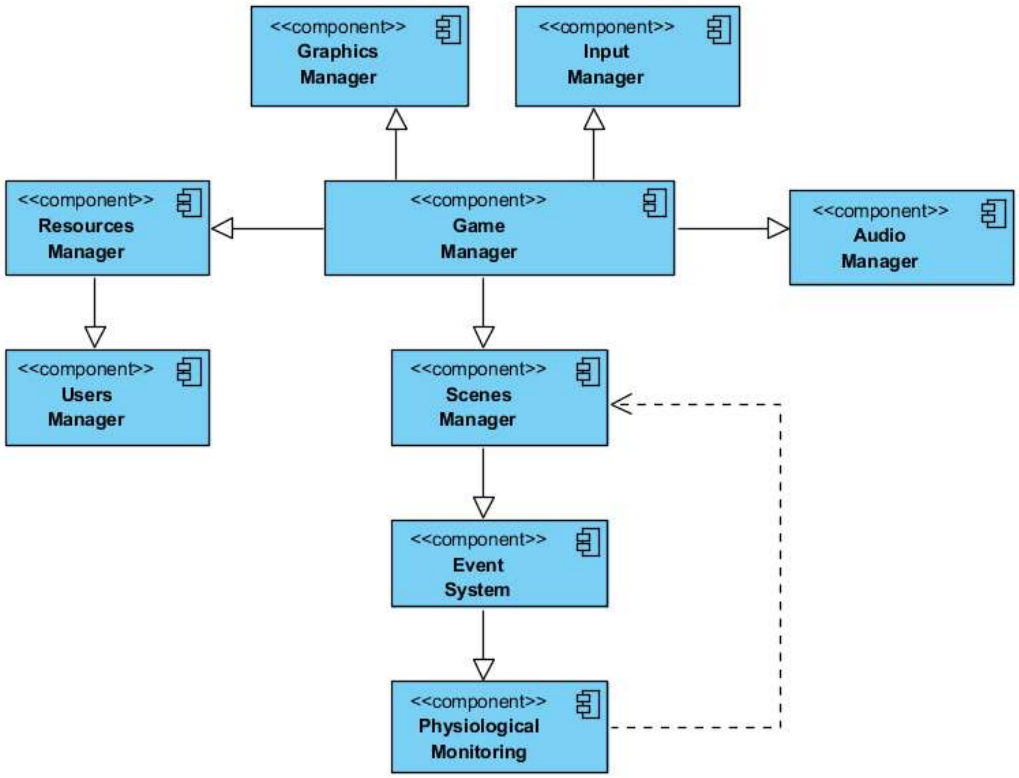

Fig. 15. Acrophobia VRET system software architecture

6.2. Development methodology

In this section we present the detailed steps of our development methodology for the acrophobia VRET system we propose.

Level I – Users analysis

At this stage, we identified the patients' and therapists' profiles, as well as what they expect from the system. The patients expect an immersive virtual environment, with a high level of realism, accessible tasks and a diverse range of in-game activities. For this, we developed the Scenes Manager module with increased attention to details in order to ensure a high level of immersion. All the 3 scenes – cable car, ski lift and walking route are carefully designed, having their graphics adapted to be rendered via the HTC Vive glasses. The input modality is also accessible and easy to be used. Thus, the patient interacts with the environment by pressing a few buttons from the Vive controller to teleport himself in the scene, select his responses to the mathematical questions and introduce the valence/arousal/dominance ratings, start, stop or exit the game, select the preferred scenario. The patients are however reluctant towards heavy and uncomfortable biophysical

equipment. Thus, even if at the beginning of our research we pursued the idea of recording EEG data, we finally dropped it out and kept only GSR and HR. These biophysical signals have been recorded using the Shimmers Multisensory device which has integrated compatibility with the C# programming language through its API [37]. The therapists expect a reliable system, with a high level and realism and immersion, visualization of what the patient is seeing in the virtual environment, as well as of his biophysical signals, so that they can easily adapt the exposure scenario. In addition, they want to have access to recordings of the users' performance in order to calculate statistics and perform post-therapy analyses. To accomplish these requests, we developed the Physiological Monitoring module and the Users Management module.

Level II – Domain analysis

At this stage of development, we interacted with psychologists and psychotherapists, in order to understand the emotional profile of the people suffering from acrophobia. Here we researched the domain of affective computing and defined *fear as an emotion with low valence, high arousal and low dominance*. The psychologists advised us to repeatedly ask the patients for their self-reported valence, arousal and dominance ratings, but before it is recommendable to detach them from the current intense emotional state, deactivate the right hemisphere and activate the left one by applying some mathematical quizzes. By listening to his favorite piece of music, look at a picture or read his preferred quotation, the patient also achieves a high state of relaxation, being at the same time an effective self-reward solution. The person rewards himself from time to time after experiencing a stressful situation or expecting to reach a certain game level, so that he can take a break, stop the exposure temporarily and enjoy a short, but pleasant activity. The data collected (biophysical signals and valence/arousal/dominance ratings) will be further used for designing an additional module, called Fear Estimation. Several machine and deep learning techniques will be used to construct a model that automatically determines the patient's current level of fear, so that the therapist will know not only the biophysical raw values, but also whether the user experiences low/medium/high fear. In this way, he can adapt the exposure scenarios more easily. Future plans include the development of an Automatic Exposure Adaptation module, where, based on the knowledge collected from the therapist, the patients' biophysical data and fear level estimation, a virtual therapist will adapt the level of exposure automatically, without or with minimum intervention from the human expert. In addition, in our future research, we intend to integrate a form of neurofeedback, so that the elements from the virtual environment would change their appearance according to the user's emotional state. So, the sky can become cloudier or darker when the patient feels anxious, clearer when he is calm and change dynamically during the session. By being provided with this form of feedback, the patient can struggle to relax and induce himself a state of relaxation in order to change the appearance of the natural elements from the virtual environment.

Level III – Task analysis

Here we identified the tasks and corresponding subtasks. The user can select at the beginning the route he wants to take – a ride by cable car, ski lift or a walk by foot. Throughout any of these routes, there are 10 stop points where a mathematical quiz is applied. After the user correctly answers the mathematical question, he is required to select his valence, arousal and dominance levels using Self-Assessment Manikins. We established the interaction between the human and the machine, communication protocols, user interfaces. All the system's tasks – patients management, virtual exposure management, physiological monitoring, application logic, flow control – have been designed and implemented at this step.

Level IV – Representational analysis

Here we established which will be the virtual scenarios, with both their graphical and audio components. Such, we designed a landscape with forests, cliffs, canyons, peaks, a cable car, a ski lift, a transparent platform and all the visual elements. As audio elements, we can mention the sound of bird chirping and the wind. At this stage of research, we have only one virtual setting, i.e. the mountain, but very shortly we will develop a cityscape with tall buildings, glass elevators, terraces and balconies.

7. Conclusions

This paper presented a human-centered methodology inspired from the HCDID model proposed by [45] and from the NADI model of Mieke and Dorst [42] for designing and developing a VRET system. The four stages of development – Users analysis, Domain analysis, Tasks analysis and Representational analysis have been adapted for the development of a VRET application dedicated to acrophobia therapy. We have carefully followed these steps and, by taking into account the patients' and therapists' requirements in a human-centered fashion, succeeded to obtain 9 functional modules responsible with users management, physiological monitoring, event triggering and audio

& graphical management. The human-centered perspective is ensured by the virtual environment's level of realism and real life inspired tasks, the first person perspective in the game that is adapted according to the player's height and by the fact that the scenario is receptive to the user's needs, so that he can relax anytime by looking at his favorite photo, listen to his favorite piece of music or read a quote he enjoys. This system of rewards is not only encouraging, but also motivating and pleasurable. We payed attention to the modality in which the user provides his emotion dimension ratings. At a psychologist suggestion, we provided a modality of deactivating the right cortical hemisphere responsible with affect and activate the left one that is responsible with thought and logic. Thus, the user is asked a mathematical quiz before introducing his emotional ratings. Also, in order to establish the mathematical skills, each user receives a test before starting the virtual reality exposure. Based on the results obtained in this test, he can receive either low difficulty / medium difficulty or high difficulty mathematical questions in the game.

Our system can collect and store data from the patients and from the therapists. This data will be used for constructing a computational model that will automatically recognize the patient's current fear level and adapt the scenario accordingly, without or with minimum intervention from the human specialist.

Future plans include performing a set of experiments with people suffering from acrophobia, collecting data and designing a computational model for emotion recognition and exposure adaptation by using various feature extraction and effective machine learning techniques.

Acknowledgement

This work has been funded by UEFISCDI proiect 1/2018, UPB CRC Research Grant 2017 and EEA-RO-2018-0496.

References

1. Acarturk, C., Smit, F., de Graaf, R., van Straten, A., ten Have, M., Cuijpers, P: Economic Costs of Social Phobia: A Population-Based Study. Journal of Affective Disorders 115, 421-429, (2009)
2. Activity, https://www.unitylab.de/. Accessed April 1, 2019
3. Arachnophobia, https://store.steampowered.com/app/485270/Arachnophobia/. Accessed April 1, 2019
4. C2Phobia, https://www.c2.care/en/c2phobia-treating-phobias-in-virtual-reality/. Accessed April 1, 2019
5. Coelho, C. M., Silva, C. F., Santos, J. A., Tichon, J., & Wallis, G.: Contrasting the effectiveness and efficiency of virtual reality and real environments in the treatment of acrophobia. Psychology Journal, 6(2), 203–216, (2008)
6. Ekman, P., Cordaro, D.: What is Meant by Calling Emotions Basic. Emotion Review, 3(4), 364–370, (2011).
7. Ekman, P., Friesen, W.V: Constants across cultures in the face and emotion. J. Personal. Soc. Psychol., 17, 124–129, (1971)
8. Ekman, P.: A methodological discussion of nonverbal behaviour. Journal of Psychology, 43, 141-149, (1957)
9. Emmelkamp, P. M., Krijn, M., Hulsbosch, A. M., de Vries, S., Schuemie, M. J., & van der Mast, C. A.: Virtual reality treatment versus exposure in vivo: A comparative evaluation in acrophobia. Behavior Research and Therapy 40, 509–516, (2002)
10. , T., Jones, P., Kasif, S.: Human centered Systems: Information, Interactivity, and Intelligence. Report, NSF, (1997).
11. Garcia-Palacios, A., Hoffman, H. G., Kwong See, S., Tsai, A., & Botella, C.: Redefining Therapeutic Success with Virtual Reality Exposure Therapy. CyberPsychology & Behavior 4(3), 341–348, (2001). Available: http://online.liebertpub.com/doi/abs/10.1089/109493101300210231
12. Gillies, M., Fiebrink, R., Tanaka, A., Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux: Human-Centered Machine Learning. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16). ACM, New York, NY, USA, 3558-3565, (2016)
13. Hockenbury, D. H., Hockenbury, S.E.: Discovering psychology. New York: Worth Publishers, (2007).
14. Izard, C.E.: The emotions and emotional constructs in personality and culture research. In R.B. Cattel (Ed.), Handbook of modern personality theory, Chicago, IL: Aldine, (1969).
15. Jaimes, A., Sebe, N., Gatica-Perez, D.: Human-centered computing: a multimedia

perspective. In Proceedings of the 14th ACM international conference on Multimedia (MM '06). ACM, New York, NY, USA, 855-864, (2006)
16. Jerald, J.: The VR book: Human-centered design for virtual reality, ACM, (2016)
17. Lang, P. J., Bradley, M. M., Cuthbert, B. N.: International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Technical Report A-6. Gainesville, FL: University of Florida, (2005)
18. Limbix, https://www.limbix.com/. Accessed April 1, 2019
19. Limelight, https://store.steampowered.com/app/544880/Limelight_VR/. Accessed April 1, 2019
20. Masood, N., Farooq, H.: Investigating EEG Patterns for Dual-Stimuli Induced Human Fear Emotional State, Sensors, (2019)
21. Mauss, I.B., Robinson, M.D.: Measures of emotion: A review. Cogn Emot., 23(2), 209–237, (2009)
22. Nation Wide Phobias Statistics, https://blog.nationwide.com/common-phobias-statistics/. Accessed April 1, 2019
23. Olesen, J.: Phobia Statistics and Surprising Facts About Our Biggest Fears (2015) http://www.fearof.net/phobia-statistics-and-surprising-facts-about-our-biggest-fears/. Accessed April 1, 2019
24. Opris, D., Pintea, S., Garcia-Palacios, A., Botella, C., Szamoskozi, S., David, D.: Virtual Reality Exposure Therapy in Anxiety Disorders: A Quantitative Meta-Analysis. Depress Anxiety. 29, 85-93, (2012)
25. Phobias Statistics, http://www.fearof.net/phobia-statistics-and-surprising-facts-about-our-biggest-fears/. Accessed April 1, 2019
26. Phobos, https://samsungvr.com/view/Uu9ME9YXR_B. Accessed April 1, 2019
27. Plutchik, R.: Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, 1, New York: Academic (1980).
28. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17(3), 715–734, (2005)
29. PSIOUS, https://www.psious.com/. Accessed April 1, 2019
30. Reitano, N.: Digital Emotions: the potential and issues of Affective Computing systems, M.Sc. Interactive Digital Media, (2018)
31. Ritchie's Plank Experience, https://www.viveport.com/apps/9347a360-c6ea-4e35-aaf1-9fab4f41cb79/Richie's_Plank_Experience/. Accessed April 1, 2019
32. Russell, J.A.: A circumplex model of affect. Journal of Personality & Social Psychology 39, 1161-1178, (1980)
33. Samsung Fearless Cityscapes, https://www.oculus.com/experiences/gear-vr/821606624632569/. Accessed April 1, 2019
34. Samsung Fearless Landscapes, https://www.oculus.com/experiences/gear-vr/1290835750988761/. Accessed April 1, 2019
35. Seffah A., Gulliksen J., Desmarais M.C.: An Introduction to Human-Centered Software Engineering. In: Seffah A., Gulliksen J., Desmarais M.C. (eds) Human-Centered Software Engineering — Integrating Usability in the Software Development Lifecycle. Human-Computer Interaction Series, vol 8. Springer, Dordrecht, (2005)
36. Seffah, A., Andreevskaia, A.: Empowering software engineers in human-centered design. In Proceedings of the 25th International Conference on Software Engineering (ICSE '03). IEEE Computer Society, Washington, DC, USA, 653-658, (2003)
37. Shimmer Sensing, http://www.shimmersensing.com/. Accessed April 1, 2019
38. Stim Response Virtual Reality, https://www.biopac.com/application/virtual-reality/. Accessed April 1, 2019
39. The Climb, http://www.theclimbgame.com/. Accessed April 1, 2019
40. Trautmann, S., Rehm, J., Wittchen, H.-U.: The Economic Costs of Mental Disorders. EMBO Rep, 7, 1245-1249, (2016)
41. Uhrig, M.K., Trautmann, N., Baumgärtner, U., Treede, R.D., Henrich, F., Hiller, W., Marschall, S.: Emotion Elicitation: A Comparison of Pictures and Films. Front. Psychol., 7:180, (2016)
42. van der Bijl-Brouwer, M., Dorst, K.: Advancing the Strategic Impact of Human-Centered Design. 0142-694X Design Studies, 53, 1-23, (2017)
43. Virtual Reality Medical Center, https://vrphobia.com/. Accessed April 1, 2019
44. Virtually Better, www.virtuallybetter.com. Accessed April 1, 2019
45. Zhang, J., Patel, V.L., Johnson, K. A., Malin, J., Smith, J.W.: Designing Human-Centered Distributed Information Systems. IEEE Intelligent Systems 17(5), 42-47, (2002).

Ethical Aspects of Automatic Emotion Recognition in Online Learning

Gabriela Moise
Petroleum-Gas University of Ploiești, Ploiești, Romania
Elena S. Nicoară
Petroleum-Gas University of Ploiești, Ploiești, Romania

Abstract

Education is deeply impacted as a result of the rise in AI technology. Classic e-learning and online learning systems do not have the ability to adapt the learning process to students' learning profile, or to their emotional state. Since emotion is relevant to learning and deeply related to both cognition and performance, automatic emotion recognition is given increasing attention in online settings. Ethics in affective computing is insufficiently represented in the literature and quasi-absent in the regulatory activity of authorities on supervising the use and development of AI technologies. Our reflective and exploratory research identifies several ethical concerns and analyses the ethical risks of automatic emotion recognition in education. Benefits and complex ethical risks are identified, and a model for responsible automatic emotion recognition systems is proposed in order to provide possible solutions to ethical issues and a specific use case.

Keywords: emotion recognition, ethical considerations, affective computing, online learning, machine learning

Introduction

Artificial intelligence (AI) is evolving at a fast pace and becomes ubiquitous in many domains, including education. It is imperative that humans raise their awareness regarding the impact of AI in everyday life and act accordingly. Due to its vital role in the society, education is to foster its strategies, span, infrastructure and skills so that it may face new challenges (EGE, 2012; COMEST, 2019). Automatic emotion recognition (AER) for the educational process comes with major ethical implications that need to be carefully considered.

Writings and theories on ethics and morality date back to ancient times; philosophers such as Socrates and Plato pioneered the field, and Aristotle was the first that used the term 'ethics'. Even if no general agreement on the use of 'morality' and 'ethics' terms was achieved, the convention that an ethical judgement has to be moral and grounded in reason proves reasonable. Emotion and whatever relates to it

are naturally connected with ethics, as emotion is the root of what makes people moral beings (Cowie, 2015).

Machine ethics focuses on assuring that the behaviour of machines in relation with humans and other machines is 'ethically acceptable', its essential goal being "to create a machine that itself follows an ideal ethical principle or set of principles" (Anderson and Anderson, 2007, p. 15). A distinction between implicit and explicit ethical machines is proposed by Moor (2006, p. 19): implicit ethical machine refers to machines coded "to promote ethical behaviour", while for an explicit ethical machine ethics are explicitly represented and the machine operates effectively based on this knowledge. The term 'ethically aligned design' in AI is mapped to design processes that explicitly include human values (IEEE, 2018) and for that purpose, data engineers educated in ethics are a prerequisite (COMEST, 2019).

Practical ethics is focused on common-sense basic principles (Ross, 1939) used to evaluate a person or action as being ethical or unethical. These principles, which could apply to any entity, are: fidelity, reparation, gratitude, non-maleficence and beneficence. Autonomy and equity were added by Peter Goldie (Goldie et al., 2011), the most famous philosopher who wrote on ethics in affective computing according to Cowie (2015).

Concerns regarding ethics in AER for online learning are manifold: Do affective computing and automatic emotion recognition have a solid scientific foundation from an ethical perspective?, To what extent ethics is considered in affective computing (AC) and AER, both in the literature and in the deployed systems?, Which are the benefits of AER in education?, and Which are the ethical risks of AER in education, including both the overlooked and the hidden ones?.

This chapter is structured as follows: an overview of AC and AER (in online learning) is presented, along with their ethical implications, in order to note the level of attention paid to ethics in AER for online learning; several ethical guidelines and frameworks in the literature are considered. Roles and benefits of the AER in online learning are identified and a comprehensive list of specific ethical risks is provided. A special section is dedicated to the ethical model for AER systems in online learning that we propose to inform the decision making factors in the development, usage and evaluation of AER-based systems. Three case studies of AER systems in education are reviewed and a use case is proposed. The final section is dedicated to discussions and conclusions, pinpointing the problems encountered in AER ethics, our achievements and future research ideas.

Emotions, affective learning and ethical implications

The process of designing, coding, and deploying AI-based systems cannot be superficial. Every stage of the lifecycle of AI-systems requires solid ethical principles and values to comply with and, therefore, it should be grounded in science psychology rather than in computing (Cowie, 2015). The novelty of the AI ethics as a research topic is witnessed by Bakiner (2022) in a survey he carried out on over 221 academic journal papers. Although there have been registered some progress in the terminology and similar

problems and solutions in different domains, the review of the existing literature has revealed a small number of commonly cited authors, Bakiner concluding that AI ethics is not a consistent field yet.

Researchers, practitioners and the public approach AI ethics differently, namely: (1) as a solution to problems where human or machine agents must make difficult decisions, (2) as a call for reflectiveness on AI technologies, (3) as an adequate distribution of moral and legal responsibility between humans and machines, or simply (4) as a critique of systems and institutions for the oppression and exploitation (Bakiner, 2022).

Rosalind Piccard (1995, 1997) pioneered the affective computing domain to identify the role of emotions in the complex human-machine relationship and to study new abilities of the machine, namely those of recognizing and expressing emotions. AC was defined as "computing that relates to, arises from, or influences emotions" (Piccard, 1995, p. 1). It may influence various fields, such as: education, healthcare, games, software development, marketing, website customization, etc.

Nevertheless, AC has profound moral implications, derived mainly from mimicking and influencing human free will, that require high attention of all stakeholders. It comes with two contradicting positions: either it can serve the betterment of the community, or it can contribute to its destruction (Cowie, 2015). Accordingly, machine ethics plays a major role in making AC serve the community (Fig. 1).

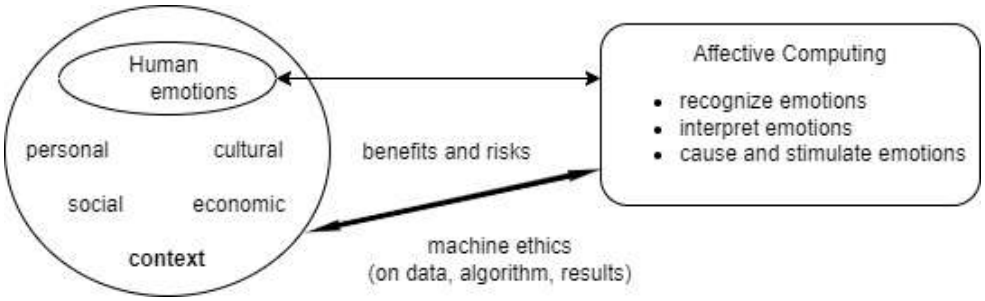

Fig. 4.1. Ethics in affective computing

To build AER systems, various models of emotions are used, mainly the discrete model centred on basic emotions (Ekman et al., 1969; Ekman, 1999) and the dimensional model based on emotion valence, arousal and dominance (Mehrabian, 1996; Russell & Mehrabian, 1977). Other emotion theories have been developed, such as the cognitive appraisal theory of emotions (Scherer, 1999) and the theory of constructed emotions (Barrett, 2017). Nevertheless, these are not considered in AER-based systems, as they are difficult to be coded in such systems. Moreover, a difference between trait emotions (affective tendencies, such as depression and anxiety) and state emotions (emotional episodes, as joy or sadness) exist. As Hascher (2010) remarks, trait emotions influence learning probably in a larger extent than state emotions.

In spite of the strong, complex, and bidirectional relationship between emotions and learning, no general well-specified rules were found and research on emotions in education was limited prior to the 1990s.

Anxiety represents an exception in this respect, as it has been a research topic since before the 1950s (Pekrun et al., 2002; Zeidner, 1998; Pekrun, 2005; Rosenfeld, 2022).

The term 'affective learning' was introduced in "Affective Learning – A Manifesto" (Picard et al., 2004) to cover all topics related to students' affective (or emotional) states in learning condition and the usage of technology in emotion recognition. Most of the research has been conducted to identify state emotions experienced by students in various learning scenarios, including technology-based contexts. A meta-analysis of 24 studies considering four such contexts (i.e. intelligent tutoring systems, serious games, simulation environments, and simple computer interfaces) found that engagement is the predominant affective state in learning (D'Mello, 2013). The importance of detecting students' engagement in the online learning in order to create an effective learning process is highlighted by Dewan et al. (2019). Other general state emotions in academic contexts are: pride of success, anxiety, hopefulness, hopelessness, relief, enjoyment of learning, anger, shame, boredom, surprise, sadness, frustration, confusion, happiness, fear, joy, disgust, interest, curiosity, contempt, delight and excitement (Pekrun et al., 2002; D'Mello, 2013; Yadegaridehkordi et al., 2019).

It has been demonstrated that positive and negative emotions improve or impede learning, respectively. An experiment conducted on 118 college students proved that comprehension and knowledge transfer increase by inducing positive emotions through specially designed learning resources (Um et al., 2012). Most positive emotions are beneficial for creativity, for example. Negative emotions, such as confusion, can be useful in learning if controlled, regulated, and resolved (D'Mello et al., 2014). However, it is naive to consider that positive emotions have exclusively positive effects on learning and that negative emotions have only negative effects on learning (Hascher & Edlinger, 2009).

A comprehensive study on AC in education (Yadegaridehkordi et al., 2019) analyses 94 papers published in prestigious databases between 2010 and 2017, and reports the following: 1) an increasing trend of AC in education, 2) the relationship between emotion, motivation, learning style, and cognition as main research topic discussed in the literature, 3) the use of visual, textual, vocal, physiological, and multimodal channels for affective measurements, and 4) the preference for using dimensional models in the education area.

Online education is, undoubtedly, experiencing a broad acceptance and rapid development, and emotions need to be suitably addressed in this context. As teachers and students do not generally communicate in a synchronous manner within online environments, teachers cannot recognize learners' emotions, so their teaching methods are not adapted to the actual learning context (Duo & Song, 2012). When the machines mediate the learning process, automatic emotion detection is envisaged to streamline the process, open datasets start to be available (Dewan et al., 2019) and AI is employed for this purpose.

Nevertheless, the ethical impact of AC on education is mildly discussed in the literature and only few academic papers contain ethics statements. They mainly regard the ethics commission' approval for the research (Cowie, 2015; Ashwin & Guddeti, 2020; Kazemitabar et al., 2021; Arya et al., 2021), users' privacy (Sharma et al., 2022; Vidanaralage et al., 2022), data protection and the informed consent if

physiological signals are used (Arya et al., 2021), the bias caused by small context-specific datasets (containing accurate expressions of affective states) or data protection for underage students (Dai & Ke, 2022).

Emotion recognition algorithms are studied from the performance perspective, but their ethical aspects and disadvantages are generally neglected. The main drawbacks of detecting emotions based on facial expressions are caused by the insufficient knowledge on expressing and perceiving emotions (Barrett et al., 2019). A more profound aspect about AI, emotions and facial recognition is that AI is portrayed by the entities that produce and profit from AI technology as a way of understanding people and the world, claiming that neural networks 'predict' personality based on facial analysis; moreover, these entities lead to the idea that the use of AI systems is common sense (Goldenfein, 2020). Online proctoring, for example, beyond defeating academic dishonesty, raises critical ethical concerns (on domestic surveillance and autonomy denial). Applications such as Proctorio, Examus and Honorlock record images, ambient sound, motion, keyboard and other device usage, screen and browser activities, and may require a smartphone for a 360° view of the student's environment. Additionally, Cowie (2015) notes the deception associated with automatic agents working with emotions being perceived as having emotional competence.

In sensitive areas such as education, the regulators play an essential role in designing ethically aligned AI. The AI Now Institute within New York University construe that regulators "should ban the use of affect recognition in important decisions that impact people's lives and access to opportunities", including those related to evaluating students' performance in school (Crawford et al., 2019, p. 6).

Ethical guidelines and frameworks

Governmental bodies, authorities, international commissions, national regulators, and professional associations responsible for regulating AI technology design and use have issued and updated ethical guidelines, frameworks, and codes, but only few have the normative role and power of enforcement over the organisations which deploy AI (Crawford et al., 2019). In UK, for example, no regulatory body is charged with supervising the use of AI and ML (IFOW, 2020). Critical issues related to this topic are:

- the difficulty of regulating AI field as a result of the impossibility to monitor it at the global level (COMEST, 2019);
- the difficulty of identifying the right decider on the reasonability of actions in an AI context (IFOW, 2020);
- the probability of rejecting AC as being ethically unacceptable by people (Cowie, 2015).

Several ethical codes may be useful for AER (i.e. APA, 2017; ACM, 2018; EUCFR, 2012; EGE, 2012), but ethics is only partially addressed, considering exclusively aspects such as privacy, access, confidentiality, integrity of personal data, explicit consent, protection against cybercrime.

Ethical values regarded as motivating ideals and foundations for ethical principles are described and structured in many frameworks (EGE, 2012; ACM, 2018; Leslie, 2019; Jobin et al., 2019; ICO, 2020;

UNESCO, 2021; Mohammad, 2022, etc.). A review of 84 frameworks and guidelines reveals the commonly addressed ethical principles, namely: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al., 2019). Transparency includes the mandatory personal data breach notification and managing one's personal data (EGE, 2012). Accessibility, autonomy, data protection, beneficence, robustness, safety and security, trust, awareness and literacy, multi-stakeholder and adaptive governance and collaboration, human oversight and determination are other ethical principles (UNESCO, 2021; Jobin et al., 2019).

A salient ethical framework for "responsible design and implementation of AI systems in the public sector", is provided by Leslie (2019). The author proposes a three-level structure for the framework and four purposes to be aimed at within AI projects: permissibility, fairness, trustworthiness, and justifiability. Due to its rigorous, comprehensive and well-structured ethical perspective in all stages of an AI project, this framework will be the foundation of our AER model further detailed.

The first level (L1), "SUM Values", is governed by the ethical values that "Support, Underwrite, and Motivate" the responsible initiative. It addresses the impacts of the project on the communities and its key incentives are: "respect the dignity of individual person; openly, sincerely, and inclusively connect with each other; care for the wellbeing of people; and protect the social values and public interest" (p. 10). The second level (L2), "FAST Track Principles", provides applicable directions to the responsible design and use of AI technology. It refers to fairness, accountability, sustainability, and transparency. Fairness settles the principle of discriminatory non-harm and data fairness (properly representative, relevant, accurate, and generalizable datasets for training and testing), design fairness (reasonable, morally non-objectionable, justifiable correlations and inferences), outcome fairness and implementation fairness (responsible deployment). Accountability addresses responsibility, and sustainability assures an ethical long-term effect of the AI system on the community. Transparency refers to the interpretability and explicability of AI systems and represents a solid justification in favour of AI-based technologies. The most practical level (L3) is the "PBG Framework" (Process-Based Governance Framework). It provides a transparency-based way to integrate SUM Values and FAST Track Principles in the AI project development workflow. Leslie's framework can be rigorously implemented provided that the project development team "Reflect" on "SUM Values", "Act" according to "FAST Track Principles", and "Justify" based on "PBG Framework", which means that the organization is driven by authentic ethical principles established by inherently moral people.

A very useful AI wide-range risk analysis, based on a 106-question template for a responsible AI project, indicates the risk levels for every adverse impact, where gravity potential and number of rights-holders affected are considered (Leslie et al., 2021).

Before developing any AC technology, several vital questions need to be addressed: Is the recognition of human emotion by machines ethical or not?, Why should such a technology be developed?, Who will benefit from it and who will not? (Mohammad, 2022), How far does ethical accountability go? (Cowie,

2015). Additionally, Leslie et al. (2021) assert that the team should not elaborate the AI project if its lawfulness is not clearly established.

Recommendations for ethical AC/AER are found in all mentioned guidelines and frameworks. Mohammad (2022) provides 57 such prescripts, including: be aware that privacy is not about secrecy, but personal choice; choose to use intrinsically interpretable / clear box models (where users can easily understand why the system predicts the result) vs. opaque box models (where users need additional tools to understand the prediction reasons); mind that all means of AER system can be misused; include neurodiverse and neurotypical participants for data annotations; be aware that no emotion recognition method is perfect. Other recommendations to follow are: do risk prevention, take mitigation and monitoring measures (UNESCO, 2021), use multimodal recognition for emotions (e.g. by facial expressions, dialogue and posture together), anonymize data and use only those that are necessary for the targeted purposes (COMEST, 2019; CDDO, 2020), responsively handle the profiles assigned to data subjects (EGE, 2012), include human-in-the-loop or human-on-the-loop mechanisms, and methods for users to be able to opt to revert to human intervention when high level of automation is present in the system (Leslie et al., 2021), employ explicitly given consent rather than assumed consent of the subjects, verify emotion identification with two or three persons in different areas of expertise and assure that the expansion of the system is controlled by certain rules.

Automatic emotion recognition in education: roles, benefits and ethical risks

Giving 'emotional abilities' to computers, as AER aims at, refers to inferring emotions felt by the user, emotions that the user attempts to convey, emotions triggered in the user, emotions' intensity, moods and emotions dynamics, attitudes and sentiments toward a target (Mohammad, 2022). Ethics in emotion perceiving by computers is still at its beginning, as ethics focuses traditionally on assessing actions rather perception (Cowie, 2015).

AER uses the capabilities enabled by computer vision and computer listening; by that, everything about the 'real world' that is visible, audible, or otherwise sensible, is recorded, computationally analyzed, and classified in real time (Goldenfein, 2020). Data are extracted from facial micro-expressions, iris data, gait, stance and gesture, speech, voice intonation, biophysical signals (skin and blood conductance, blood flow, respiration, infrared emanations, brain waves), haptic data such as force of touch, typed text, emoticons, emojis or self-reported questionnaires.

Various machine learning (ML) techniques are used. Despite their results, major drawbacks regard unrepresented people and the fact that huge balanced data are needed for a better representation. Nevertheless, one should be aware of the fact that the more data, the higher ethical issues and risks.

To identify the ethical impact of AER-based online learning, we first reviewed the roles of AER systems in online learning. ScienceDirect database returns 102 results regarding roles of AER systems in online learning for the interval 2015 – 2022, using "emotions recognition" and "online learning" as search

phrases. Relevant sources were determined based on the titles and by abstract/preface screening, and finally a full text screening was performed. Out of the 102 scientific works, 27 (26.5%) were considered relevant for our purpose, 26 studies referring to students' emotions and only one to teachers' emotions. Only five of the 27 (18.5%) include considerations on ethics.

The list of AER systems roles with their benefits was supplemented with the results of our academic experience and presented in Table 1. For each role, both students and teachers are envisaged as possible users.

Table 1. AER roles and benefits in online learning

AER role	AER benefits	Resources
Designing and building Intelligent Tutoring Systems (ITS) as pedagogical agents with emotional abilities, which adapt instructions to students' performance and learning profile (based on the relationship between students' emotions, motivation, cognition, learning styles)	• customized feedback; • better effectiveness of learning; • support for teachers regarding educational resources and teaching strategies;	Xu et al., 2018; Alwadei&Alnanih, 2022; Kazemitabar et al., 2021; Sikström et al., 2022; Dai&Ke, 2022; Cen et al., 2016; Chen&Wu, 2015; Feidakis, 2016; Faria et al., 2017; Lin&Kao, 2018; Yang et al., 2018; Imani&Montazer, 2019; Ez-zaouia et al., 2020; Iulamanova et al., 2021
Supporting engagement and motivation of learners and teachers	• better engagement and motivation of the participants; • support for raising the students' awareness on their mental and emotional states;	Chen&Wu, 2015; Feidakis, 2016; Yadegaridehkordi et al., 2019; Hasnine et al., 2021; Bhardwaj et al., 2021; Lavoué et al., 2021; Liu et al., 2022; Sharma et al., 2022; Qiao et al., 2022; Vidanaralage et al., 2022; Lyu et al., 2022
Learning assessment	• fraud-free results;	Tanko et al., 2022
Teaching assessment	• support for raising the teachers' awareness on the role of emotions in teaching;	Utami et al., 2019
Building comfortable learning environments	• more effective learning and teaching;	Arya et al., 2021; Alfoudari et al., 2021
Supporting students with special needs (ADHD, anxiety and so on)	• better educational help for students with special needs.	Alwadei&Alnanih, 2022; Sikström et al., 2022

To identify all the ethical risks or actual harms, an iterative search-review-discussion process was carried out. There was generated a preliminary list with 24 risks identified both in the literature and in our academic experience. Based on the authors' experience of 20+ years in computer science and artificial intelligence, these risks were systematized in sixteen classes. As the study focuses mainly on students, a brainstorming session with ten computer science students followed in order to refine the preliminary list and teachers' observations, and a final list with possible ethical risks was obtained.

The targeted AER ethical risks reported in the literature are manifold. Some are general AI risks (as presented by Leslie, 2019; Cowie, 2015; EGE, 2012; Goldenfein, 2020; COMEST, 2019; Sadowski, 2021), whereas others are risks specific to the educational process (Lyu et al., 2022; Mohammad, 2022; Arya et al., 2021; Sharma et al., 2022). Some of the sixteen identified classes of AER's ethical risks in education have many facets and more practical implications, whereas others have inherently descriptive titles. The sixteen classes are:

1. Bias and discrimination, due to poor representativeness of data or based on the designers' preconceptions or bad intent. Only large volumes of balanced representative data are adequate for proper results, but the independent rigorous researchers generally have a limited access to data which technology companies own (Sadowski, 2021). Moreover, there is enormous variability in human mental representation and expression of emotions (Mohammad, 2022). AI is "not neutral, but inherently biased", a reason being that "classification is culture-specific and a product of history" (COMEST, 2019, p. 8, 7). For example, recommender agents suggest items by using discriminating filters (Shelton, 2022) and, as a result, certain resources are presented to students, whereas others are hidden, based on the ML designers' line of thought.
2. Unreliable, unsure, unsafe or poor results. AER systems may indicate causal relations between data, which in fact do not exist (e.g. racial or gender differences in intelligence or learning outcomes) and could poorly recognize emotions by ignoring for example the possible distracting factors in the environment. Consequently, some inferred emotions may be further misinterpreted.
3. Non-transparent, unexplainable, unjustifiable or not fully predictable outcomes. AI results cannot be fully predictable or explainable.
4. Privacy invasion by (1) inaccurate ownership and management of personal data, (2) failure in giving and withdrawing consent and (3) domestic surveillance. Video analytics used in proctoring or keeping attendance in class leads to vast databases with personal data. Facial expression, voice, gait, physiological signals and other biometric data are highly sensitive data. Other recorded data might include institutional affiliation of the users, their skin colour, age, gender, what he/she is doing at that moment, who the person has been associating with (Goldenfein, 2020). When interacting with emotion or facial recognition in different settings (homes, schools, or outdoors), individuals perceive their privacy differently (Sharma et al.,

2022). They often make their data available without being aware of the implicit or explicit acceptance of the hidden purposes (EGE, 2012). The risk of defective ownership and management of personal data raises when the data become accessible to many parties who do not intend to protect the subjects (Arya et al., 2021), especially when the risk of deanonymizing through data linkage with existing data is present (Leslie et al., 2021). A free and fully-informed consent of the users to participate with data is needed when using AER systems (for training or deployment, e.g. in online assessing). It includes mechanisms to ensure that the user clearly understands which data are gathered and processed, to whom they are accessible, for how long, for which goals and what are the related risks. As AER attempts to predict emotions, behaviours and personality type (Goldenfein, 2020), the risks of privacy invasion increase. Moreover, in education group privacy should be addressed.

5. Unfairness and digital division. The sequent inequality, exclusion, threat to cultural diversity and exploitation of the vulnerable groups impede education as a public service. For the students with special needs (e.g. ADHD, anxiety, depression, alexithymia, and autism), fairness and equality while handling online learning with AER is hazardous. More personal data are regularly captured from these subjects, which may lead to more privacy risks. If academic institutions used AER mandatorily, education would become accessible only to those who accept the related risks and economically afford it. The so-called "digital divide" regards the access to data, to algorithms, to human and computational resources (COMEST, 2019).
6. Deception. Unthought of or deliberate deception regarding AER are three-folded. First of all, the pseudo-scientific base of emotion recognition may mislead towards the belief that ML could ever infer one's true emotional state (Mohammad, 2022). AER can infer some aspects about one's state emotion, but this does not equal trait emotions. Variability of emotion expression, if not broadly considered, leads to the false message that AER decides which emotion is 'standard', other forms of expression being ignored or invalidated. Moreover, certain theories on emotions are simply neglected because they are difficult to be coded in data, and various factors with impact on the affect are omitted when labelling emotions (i.e physical and mental illnesses). Consequent risks for students are a limited or wrongly-conditioned access to resources, be them educational or studentship grants, and a negative influence on students' evaluation. Secondly, the illusion of 'emotions' and 'intentions' of automated systems as artefacts, in contrast with humans as responsible agents, is better to be acknowledged. On the one hand, the entire AC domain is inherently deceptive, on the other hand, people themselves cannot be labelled as dishonest when someone shows an emotion not corresponding to the internal feeling (Cowie, 2015). Thirdly, when untrue ethics is presented as true, deceiving by ethics washing occurs. Examples in online learning are a) partially-informed consent presented as an explicit and fully-informed consent, and b) avoidance of clear indication of grey area

decision making or the low number of accurate metrics. Deception becomes obviously unethical when it impairs the users' possibility to exercise autonomy (Cowie, 2015).

7. Manipulation and building authoritarian relations. This risk manifests when inferred emotions are misused. Inspecting the moments when students and teachers are most receptive for outer suggestions, their behaviour can be easily manipulated. Automate agents that would combine impeccable logic with infinite patience, no conscience, and the ability to manipulate emotion would create an almost irresistible persuader (Guerini & Stock, 2005). Also, unethical teachers or faculty could build authoritarian relations by misusing students' data in exams or in any other life event.
8. Changes in human perception of reality, understanding, expertise and natural behaviour. The term 'emotional perception' related to AER, which is far less accurate than the emotional perception of vigilant humans, could influence users' trust in their own ability to naturally perceive emotions, be them of their own or of other people. If a transparent AER-based online assessment agent explains to the student its prediction, that he/she cheated, because the student's eyes moved for one minute in the upper left part of his/her visual field, while the face turned pink, that student could wrongly change the perception regarding other people's similar actions as indicating a form of cheating. By contrast, many times, people tend to look in the upper left when they try to remember something, or to maintain their point of view while speak directly to someone; face turning pink could mean a burst of emotion of various type (shame, joy, anger etc.) or even only a physiological blood-circulation alteration. More important, the perceived meaning of 'facts' and 'truth' can be deeply influenced by AI algorithms (COMEST, 2019), which is critical for education. The more accustomed to automated decisions one is, the more their visual understanding of the world is challenged and tends to be changed, minimalized, and "even negated by computational ways of sensing and knowing" (Goldenfein, 2020, p. 5). As virtual reality and augmented reality frequently complement AER, this risk increases significantly.
9. Erroneous portraying of human beings and emotions. When people describe an emotion, they pass a moral judgement on its justifiability for example, in order to have an accurate image of that person. Is a machine entitled to do the same or can it do so? When a student is angry because a colleague is disturbing his/her learning, the AER software eventually labels anger and may connect it strictly with learning, ignoring external factors.
10. Denial or bypassing of individual autonomy and rights (restriction on users' ability to exercise free will or free speech, non-free and non-informed decisions regarding users, denial of right against self-incrimination). Users' autonomy and fundamental rights venture to be affected if their religion, lifestyle, culture and government are not envisaged while designing the AER system. Some examples of denial and bypassing risks are students or teachers may not want

their emotions to be inferred; teachers' pedagogical experience could be ignored by the automated system in evaluating students' learning profile, in using the most appropriate methods and resources for teaching; students' own educational goals may be ignored in the tutoring process. When people feel monitored, their reactions, thinking and creativity devolve. Moreover, some assessment strategies could limit the liberty of expression or students' critical-thinking. Another limitation in autonomy arises with the excessive decisional help provided by the AER system, which leads to non-versatile students, poorly resilient or lacking critical thinking, and, as a result, unprepared for the real world.

11. Dual use (the risk of using AER functionalities developed for a certain context in other more sensitive contexts, such as healthcare, civil liberties, universities or countries where data protection and other rights are not entirely observed).
12. Isolation of individuals, disintegration of social connections and dehumanizing of people relations by emotional and social interaction with high-performance, yet lacking self-awareness, AI systems. Most of the online classes deplete students' inherent ability to collaborate and clearly dehumanize human relations.
13. Dependence on a machine. Robots with 'emotional intelligence' could shape undesirable attachment of users on a machine for learning / teaching efficacy or for their emotional well-being. Emotional aid used for a long time or assisting children reduces the natural ability to self-regulate the emotional status.
14. Risk of losing the sense of individual identity. The user of an AI system loses the sense of individual identity, if the system places them on the position of an insignificant or helpless actor (Leslie et al., 2021) and even the minimization of the emotion's role may easily occur.
15. Replacement of the teachers. "One of the main societal concerns regarding AI is labour displacement", as World Commission on the Ethics of Scientific Knowledge and Technology asserts (COMEST, 2019, p. 9). With growing technology and more and more functionalities in automated tutoring, the need of continuous IT upskilling for teachers and students significantly narrows the teaching role of humans in education. The lack of transparency on this issue intensifies the job replacement hazard, while human-to-human teaching remains the main factor to foster learning (by the complex human direct interactions).
16. Lack of energetic sustainability. Ever-increasing big data pre-trained AI models require huge energy consumption and solutions with short-term efficiency.

Once acknowledged, specific risks must be properly computationally codified in order to be useful in an ethical design and deployment of AER system, as ignoring not yet encoded aspects is a risky strategy in itself (Cowie, 2015).

A critical reflection on AER facilitating a responsible emotion research and on the proper use of AER technology is provided by Mohammad (2022). The author points out the commercial and governmental

uses of emotion recognition and insists on the active engagement of the AER community in considering ethical ramifications of their creation.

Our model that will be further detailed accounts the ethical purposes in both AI and AER and proposes solutions to prevent and mitigate the negative ethical impact of these technologies.

Ethical automatic emotion recognition model for online learning

The ethical guide composed by Leslie at Alan Turing Institute (2019) represents a real tenet for the responsible AI systems. In this section, we propose a human-centered ethical AER model for online learning, based on Leslie's guide, on the Ethics sheet, on the sentiment analysis carried out by Mohammad at the National Research Council in Canada (2022), and on the data ethics framework of the Central Digital and Data Office (CDDO, 2020). In our opinion, it is mandatory that the evaluation of the AER system's feasibility be rooted in ethics and safety.

The numerous choices to make when developing and using AER in education have high long-term ethical impact. Therefore, a systemic choice architecture supporting all stakeholders (designers, testers, implementers, users, and so forth) to make their best choices (Kulkarni, 2022) is required. Various scenarios on choosing the emotion representation model, human values to base on, solutions to mitigate the tension between human variability and machine normativeness, criteria regarding people behind the data and even the working team are carefully designed and tested.

Our model (Fig. 2) is to be best implemented following Agile-Waterfall hybrid methodology, namely to design, plan and define requirements with Waterfall, and to develop and test with Agile. The highest ethical risks were emphasized for each stage of the process.

In the problem formulation stage, the goal of the proposed AER system is set, and the particular context of its application is defined, including the domain's regulatory environment, human and technological systems planned to be replaced by it (Leslie et al., 2021). The users' needs (starting with the most disadvantaged individuals in the context) and the domain-specific needs were identified; learning analytics was taken into account (i.e. an online tutoring system in Ancient History for everyone has needs that are obviously different from an online programme for a Mathematics lecture at the university); the outcomes were defined based on certain human values, objectives and beliefs; the functional design was specified (i.e. type of application, domain or use case specificity, explanatory strategies regarding the model and the outcomes); the overall impact and all potential risks were cautiously analyzed in order to separate the tolerable ones from the broadly acceptable risks and the inacceptable ones; the widest timescale in which the system could impact users was also determined (Leslie et al., 2021). Moreover, the user stories were carefully chosen and described. A solid knowledge of the field plays a major role in the impact analysis and in the final efficiency of the AER system. Moreover, prior training for implementers and users will be an advantage (Leslie, 2019).

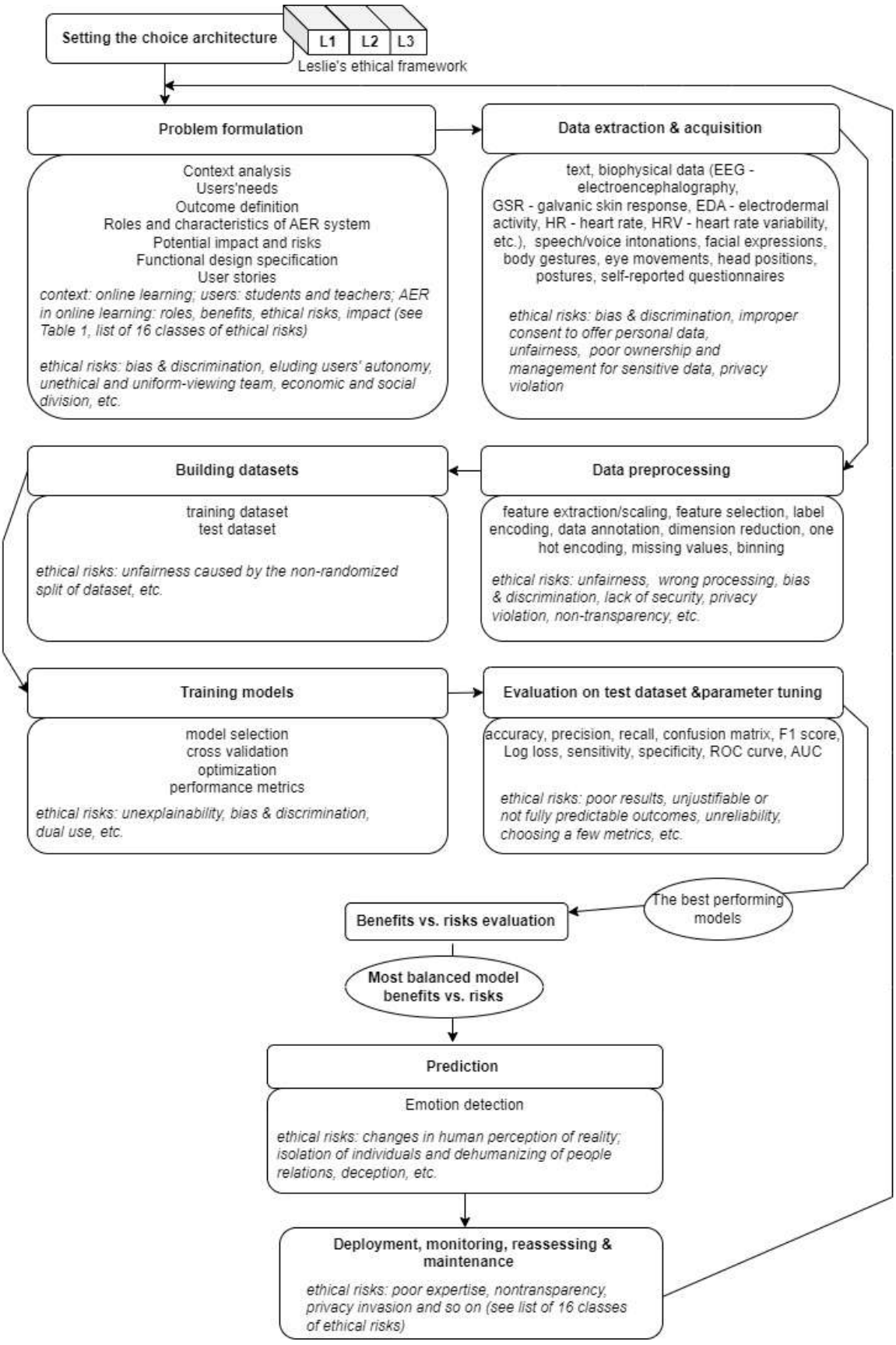

Fig. 2. Ethical prediction model for AER in online learning

For ethical data extraction and acquisition, aspects such as responsibly choosing data to be collected, finding data from diverse sets of relevant sources, and providing details about the sources are critical (Mohammad, 2022). Collecting data is the most time-consuming step in the development of a ML model, as a small number of instances manually annotated for emotions in AER systems is available. This represents the first contact that students and teachers alike have with the impact of highly sensitive data. Ethical issues may arise from the very beginning, when subjects are more or less properly informed about the real purpose of the system. As data acquisition for AER is deeply connected with experiments on human subjects, special care must be given to data integrity, free and fully informed consent, as well as to subjects' privacy.

Data preprocessing phase includes: feature extraction/scaling, feature selection, label encoding, data annotation, dimension reduction, one hot encoding, missing values and binning. Ethical issues posed in this stage point to the variability of emotion expression and mental representation, to the tendency to capture the attitudes of the majority group, to the difficulty of choosing the right perspective on what is appropriate and what is not (Mohammad, 2022), to bias and discrimination associated with crowdsourcing, and to feature omission. For automated labelling or annotation, ethical solutions are human oversight (Leslie et al., 2021) and considering multiple answers that are more appropriate than others instead of a single 'correct' answer.

With respect to the stage of datasets building, the randomized split of the data must be ensured.

For training a predictive model, the most ethically susceptible steps are the model selection and the criteria selection. In our opinion, it is recommended to choose multiple, various and proper metrics, such as accuracy, precision, recall, specificity, F1 score, degree of inappropriate biases accepted, efficiency, privacy preserving capacity, transparency, interpretability, explainability, etc. Professional and institutional transparency, which covers integrity, honesty, sincerity, neutrality, objectivity and impartiality (Leslie, 2019) must be also addressed here. Additionally, we must consider the dynamics of the individuals' emotion, i.e. the change with time of their perceptions, emotions, and behaviour.

As for model selection, data available are important. AER systems regularly use ML techniques for large datasets. AI models tend to work nicely for people well-represented in the data, but abnormally for the others (Mohammad, 2022). A preferable policy is to not use intrinsically interpretable models unless the "potential impacts and risks have been thoroughly considered in advance" and the semantic explainability has the potential to soften the potential risks (Leslie, 2019, p. 46). In the sensitive contexts (such as online AER for fraud-free exams), where the transparency is important, the interpretable ML techniques are preferable choices (i.e. linear regression, logistic regression, decisions trees, or case-based reasoning). The system rationale must be non-opaque and accessible to all affected parties' understanding, in terms of their capacity and limitations of cognition. When these interpretable techniques are inappropriate for our goal, more complex and model-specific or model-agnostic mechanisms will be used for the interpretability and explainability purpose. Technological maturity of the system will be proved, provided that the design is

based on well-understood techniques already in operation and externally validated for a similar context (Leslie et al., 2021). An efficient means to evaluate a model is to test it on unseen data using multiple various metrics. This stage results in several ML models, after which they are ethically analyzed in the benefits vs. risks analysis phase. It is unrealistic to state that there is no risk. We must be aware of all the risks, accept them and try to mitigate them if the benefits of using AER in online learning are relevant. High risks, combined with low benefits should lead us to drop out the development or usage of such a system. The most balanced model based on the benefits - risks ratio is then selected for usage.

In the prediction phase, the model outputs a result which is further used in the online learning system. The results may be emotions labels, if discrete models for emotions are used, or values in case of dimensional models for emotions. Ethical concerns regard the impact of the detected emotions on the users, as detailed in the list of risks. The detected emotion is used by the system to configure the learning environment, to provide support and so on.

A responsible implementation of the system must be followed by responsible deployment, monitoring, reassessing and maintenance, as the lifecycle of the system offers a social meaning to our initiative. These four steps must be thoughtfully approached by acknowledging the proper roles in the team and ethically professing them, so that to deliver the project in correlation with the real needs of the users, which must be constantly revisited throughout the entire process. Students and teachers have to be fully informed on the AER technologies used in the programme (either experimental or not), prior to their participation in the corresponding tasks. Their consent is to be clear and explicit. Both the implementation and the subsequent stages will benefit from training the implementers and the users, so that one may prevent biases and deliver an interpretable and justifiable system. Taking account of the breadth and temporality of deployment, of directly and indirectly affected users (Leslie et al., 2021) and of explanatory strategies will also facilitate the process. A participatory AER system, where all users are invited to make comments and offer recommendations that may improve the system, empowers them and lowers possible tensions. For a broader perspective on the student's learning results, assessment and other issues, the teacher will be the final decision-maker.

All the components of the model imply ethics in many ways and, as a result, the attention paid to the three levels (L1, L2, L3) ensures fairness, trustworthiness, justifiability, and permissibility. SUM values pinpoint the respect for students and teachers, open and inclusive connection, the wellbeing of users and the protection of social and cultural values, as well as for the public interest. FAST Track Principles followed throughout the process set fair, accountable, sustainable, and transparent directions of action, and the PBG Framework concretely integrates all these values and principles in the action.

It is advisable to investigate actual solutions, considering all identified ethical risks and answering the questions detailed in the Data Ethics Framework (CDDO, 2020). The whole process must be regularly revisited throughout the project, especially when data collection, storage, analysis or sharing is affected by any type of change (ICO, 2023). Constant feedback implies asking the team several questions, namely: in

the initial phase if "they are doing the right thing?", during the project if "they have designed it well?" and after the project being deployed if "it is still doing the right thing?" (CDDO, 2020, p. 18).

Case studies and use case

No ethically reliable academic or commercial AER system for learning has been developed so far, to our knowledge. Three ITSs in AER environments, used in the few identified experimental studies, were selected to investigate to what extent ethical issues are addressed. MetaTutor and iTalk2Learn are addressed in the only two papers, out of the 53 works analyzing students' emotions, that were selected in an excellent review of the studies about ITSs published in seven prestigious databases, namely Web of Science, PubMed, ProQuest, Scopus, Google scholar, Embase, and Cochrane (Mousavinasab et al., 2021).

MetaTutor (Azevedo et al., 2011) was used in an experimental study on emotion detection in learning, carried out by Harley et al. (2015), where 67 students were involved. It contains four pedagogical agents to facilitate self-regulated learning and employs the facial recognition software FaceReader 5.0 and the electrodermal activity data acquisiton software Affectiva's Q-sensor 2.0. Even if all the emotion detection methods used (automatic facial expression recognition, electrodermal activity and self-report) generate high-sensitive data, no ethical acknowledgement was held.

The ITS component of iTalk2Learn platform was evaluated in an experiment on the adaptation of the feedback given to the students to their state emotion. It uses multimodal emotion detection, namely speech analysis, as well as the analysis of the changes occurring in action after certain indications have been given. A possible positive role of emotion-aware technological support was identified, but, nevertheless ethical aspects were not approached (Grawemeyer et al., 2016).

Affective AutoTutor, employed in over twenty controlled experiments, uses multimodal affect detection (i.e. facial features, conversation and body language) to keep the student in a balanced emotional state, by varying the difficulty of the tasks, the pace and direction of learning (D'Mello & Graesser, 2012). It fulfills ten complex functions, whereof modelling students' cognitive states and regulating negative affective states, but, once again, the ethical aspect was neglected.

Our proposed use case for a tutoring system in the online Optimization Algorithms course, augmented with examination tools, was designed based on the prediction model described in Fig. 2. Ethical recommendations previously mentioned are to be carefully envisaged in every aspect of the use case. Such use cases will be carried out in experimental studies, where an adequate AER software for online learning will be developed.

The suggested work scenario for the use case starts by setting the choice architecture, followed by the formulation of the problem, which was achieved by setting the items:

- Team characteristics: multidisciplinary expert team with diversity of thought and wide ranging skill sets (strong ethics; solid knowledge in learning emotion-based profiling and assessment; serious analytical skills; good collaboration with instructors in computer science optimization);

the ethicists, academics, data scientists, policy experts, researchers and practitioners that clearly understand the needs of the users;

- Context: superior education, Bachelor's programme in Computer Science, online learning, Optimization algorithms course;
- Users: students and teachers;
- Needs: for students - customized learning routes in order to gain good abilities in optimization algorithms; for teachers - facilitated teaching by accounting students' emotions in the online environment;
- Roles of AER system: students' learning profiling, adaptive teaching and learning, objective fraud-free assessment; characteristics: video, sound and text recording for users' online activity, ITS with capabilities of running complex optimization applications, comparing tools for algorithms, visualisations; textual, vocal (conversation in natural language) and multimedia response for the users;
- Outcome: effective personalized teaching and learning; accurate, objective and fraud-free examination;
- Impact analysis: transformative and long-term effects on students' learning strategies and knowledge in the area, on emotion-related aspects in users' lives, on teaching strategies of instructors, and student-teacher relationship;
- Data used: text, speech, nonverbal and paraverbal communication, i.e voice intonation, facial expressions, eye movements, head position, posture, gestures, as well as self-reported questionnaires;
- User stories with specific actions mitigating the corresponding high ethical risks (see risks classes and examples in Table 2).

Table 2. User stories and ethical actions of an AER system with tutoring and examination role

User story A As a student with low level knowledge and interest in mathematics, I want general-scope training in optimization, so that I will pass the exam, being objectively assessed.	
Suggested actions	**Ethical risks**
Detect student's emotional state (confusion/ frustration/ shame/ boredom/ surprise/ hopefulness/ contempt etc.) and level of demotivation	1, 2, 3, 4, 5, 9, 10, 11, 16
Detect student's learning profile in repeated interactions (emotions – motivation - learning style - cognition)	1, 2, 3, 4, 5, 6, 9, 10, 11, 16
Recommend educational resources to the teacher (vital short videos and readings, optimization case studies in the real-world), teaching strategies adjustments (e.g.	1, 2, 3, 5, 7, 13, 14, 15

practical engaging activities, form-based tasks, clarifying explanations, semi-weekly small tasks and assignments, team working with classmates)	
Regulate negative emotional states as frustration and boredom by delivering hopeful, motivating or congratulation messages, short videos of similar successful projects, recreational video, images, animations or music to re-engage the participant	1, 2, 3, 5, 7, 13
Provide adaptive feedback to the student: support files for homework, clear explanations of their own emotions labels and of the exam result	1, 2, 3, 5, 7, 8, 12, 13
Provide feedback to the teacher: the interpretation of the student's emotions and emotion changes in the student, the impact of the teaching strategy and of the assisted assessment	1, 2, 3, 5, 7, 8, 12, 13

User story B

As a highly- motivated student, I expect to obtain the best educational and challenging practical tasks, so that my abilities in optimization will be competitive.

Suggested actions	**Ethical risks**
Detect student's emotional state (curiosity/ engagement/ surprise/ excitement/ general interest or interest in certain subareas/ anxiety/ relief/ contempt etc.)	1, 2, 3, 4, 5, 9, 10, 11, 16
Detect student's learning profile in repeated interactions	1, 2, 3, 4, 5, 6, 9, 10, 11, 16
Recommend educational resources to the teacher (short or medium-size videos, textbooks, various optimization case studies in the real-world, optimization tools demos), teaching strategies adjustments (e.g. weekly tasks and challenging homework requirements, team working or contest-like assignments)	1, 2, 3, 5, 7, 13, 14, 15
Provide adaptive feedback to the student: further readings on interesting subtopics in certain areas, greetings, access to more complex optimization tools, engagement techniques, if deadline is short etc.	1, 2, 3, 5, 7, 8, 12, 13
Provide feedback to the teacher: student's responses to the hints, student's emotion interpretation and the emotion changes in the student	1, 2, 3, 5, 7, 8, 12, 13

User story C

As an instructor, I expect fraud-free results in exams, so that assessing accuracy is maximized.

Suggested actions	**Ethical risks**
Detect student's emotional state during the exam (anxiety/ relief/ contempt/ confusion/ frustration/ fear/ hopelessness etc.)	1, 2, 3, 4, 5, 7, 9, 10, 11, 16

Send items clarification, motivating or relaxing posts, notice about the deadline, notifying messages in case of cheating detection, explanations about inferred results	1, 2, 3, 5, 7, 15
Assist the teacher in the examination process by monitoring student's activity, and by indicating any case of cheating	1, 2, 3, 5, 7, 8

In all phases of the process, there are applied measures to prevent or to reduce ethical risks by: obtaining users' informed consent for personal data collection and emotions inferring; sharing the understanding of the user's need with the user, accounting students' own educational goals; identifying and including the students in greater need or the disadvantaged ones and identifying measures to help them; reducing economic, social, gender, racial, and other inequalities; identifying users who may face negative consequences when using the system; using large volumes of balanced representative and proportional data (from people with different backgrounds) and manual checking for data labelling to avoid bias and discrimination; understanding how data are generated; ensuring data integrity; properly using synthetic data, if necessary; mitigating possible bias; using recommender algorithms with non-discriminatory filters; checking data limitations; spotting reliable and safe patterns in data; adequately processing data and feature omission; choosing an interpretable model, responsively using data anonymization; randomly splitting the dataset; designing and using explanatory strategies for the outcomes as well as various metrics to evaluate the model on the test dataset; avoiding emotions misinterpretations and misusing; using multimodal channels to interpret emotions; considering the variability in the expression of emotions, emotion dynamics, casual physical illnesses which may alter emotions; taking measures for transparency and explainability (about the system, algorithms, outcomes); respecting human rights; using GDPR (2016); including privacy by design; ensuring non-discrimination and reliable results; using mechanisms to prevent unfairness; ensuring accountability (robust practices, solid documentation, validated and reproducible algorithms); involving teacher oversight for important decisions, the use of teachers' pedagogical experience by participatory mechanisms; identifying signs of authoritarian relations and taking actions to eliminate them; fostering students' liberty of expression, critical-thinking and creativity, especially under examination; providing well-directed and non-controlled access to educational resources; sending clear message to users that the inferred emotions are only indicative and perhaps not their true emotional state; avoiding over-helping for students and their dependence on a machine; challenging students to override learning obstacles and to collaborate with teachers and other students; avoiding dual use of the system; complying with the law and additional ethical regulations; continuously evaluating the project (e.g. by self-assessment with scores for every specific action during the process); repeatedly revisiting the needs of the users; consulting the target audience on appropriateness of emotion recognition; involving experts and consultants in reviewing and assessing ethical considerations of the project; setting the end mechanism, if

the project stops being ethical; training users and the team in ethics; deploying, monitoring and reassessing the system in a responsible way and by performing maintenance for sustainability (green AI).

Discussion and conclusions

In our opinion, as students and teachers experience various emotions throughout the learning process, AER systems used in online learning in order to recognize, interpret, cause and stimulate emotions, bring several benefits in accordance with the diverse roles existing in education. Alongside these benefits, we must fully acknowledge the risks, especially the ethical ones, in their span, timescale and depth of impact. Even if the AI domain was set in the 1950's, AI ethics is still a novelty. In the literature, machine ethics is approached on four lines of thought, as we have highlighted. To separate legitimate worries of overstated ones, the ethical risks of using AI, AER and AC were analysed in this chapter, based on the exploratory research of academic publications, on the assessment of above-mentioned experimental applications and on critiques addressed by the general public.

The risks associated with the ubiquitous AI-based services and big data society represent a major concern for most of the researchers and for the general public, as well. We synthesized and structured ethical risks encountered in online learning in sixteen classes and provided scenarios of their emergence and real-world examples. These critical risks of using AER, if not addressed and mitigated, may easily lead to misdirected learning, along with all specific short or long-term consequences.

If emotion recognition by machines is ethical or not remains questionable, as it is not possible to capture one's full emotional experience, even if all possible data are collected. On the other hand, commercial AER applications are advertised as being able to detect the true emotional state of a person and to even 'predict' behaviour, mood and type of personality.

For a specific AER system, we must thoroughly weigh up both the impact and the risks. As a consequence, high risks associated with low benefits of an AER system should lead to the decision of stopping its development or usage.

The AC ethical impact (on education) has been addressed in the literature only in recent years. Noticeable research in the field started after 2010, accelerated after 2019, and it mainly regards general ethical aspects on AI and data protection. Three AER applications used in education that were employed in the few experimental studies reported in prestigious databases, were analysed and no ethical considerations were identified.

Nevertheless, many ethical guidelines and frameworks for AI projects have been issued by various governmental bodies, authorities, commissions, and independent researchers. As all the guidelines stipulate, AI-powered technologies should envisage fairness, trustworthiness, permissibility, and justifiability. These publicly available ethics guidelines, some of them containing wide-scope comprehensive checklists for AI projects and risk analysis templates, together with the research in the area, reveal an authentic propitious pursuit of moral values in the AI and AC research. They constitute in

adequate solutions and recommendations for each AI/AC technology, that, howbeit, has a significant direct or indirect impact on individuals, communities and overall society by means of private and public services (including education).

With their support, individuals and institutions alike raise their level of awareness regarding ethical AI. Nevertheless, the main critical issue arises, as the three aspects mentioned below make the ethical sustained effort either nearly fruitless, when regarding the actual deployment of such technologies, or inapplicable in nowadays society:

(1) AI technology companies continue to invest enormous resources in expedite deployment and in selling more and more advanced technologies. They own vast collections of data for ML algorithms, but data sources and the means employed in data collection or in creating algorithms still remain opaque for the public or indicate rather unethical procedures. When such technologies are promoted, AI neutrality and ethical risks are eluded and only the benefits for users are presented. Therefore, such misinformed users support the spreading of unethical or partially ethical AI technologies. For details on the manipulation of users' trust and deception, see our list of ethical risks.

(2) There operates no regulatory and enforcement authority on supervising the use and development of AI (to our knowledge). Therefore, how can one conciliate this lack with the AI guidelines stipulating for mandatory lawfulness of AI projects' design and development? Furthermore, from what position can one individual or organisation decide what is the right thing to do in a particular context?

(3) As companies and AI are ubiquitously networked, it is fairly difficult to regulate AI development both at the national and the international level.

This high-profile controversy between ethics concerns and solutions in ethical guidelines on one hand, and companies developing and promoting AI without complying to such guidelines, on the other hand, clearly indicates that AI ethics represents just a step in the technology regulation, an important one that is to be followed by further efforts.

In this chapter, a scalable ethical AER model for online learning is proposed, and specific ethical risks organized in sixteen classes of risks identified for AER in online learning are presented. In order to prevent or to mitigate ethical harms, we recommend the following actions: use strong ethical teams for systems design and development, do ethics education for data engineers, and maintain autonomy in human hands. It is advisable that AER technology remain just a support, not a determinant, for both teachers and students. For a more practical approach of the model, an use case comprising three user stories is detailed, focusing on specific ethical risks classes and implicitly on corresponding solutions.

To conclude, the aim of the present chapter is to prove that AC needs to undertake measures from the ethics perspective and that ethics must be more widely covered both in the literature and in the deployed

AER systems. Even if AER in education brings numerous benefits, there are also various ethical risks that must be addressed in order to reach the highest potential of emotion recognition systems.

For future research we suggest the following directions: (a) to describe the correlations between AER benefits, AER implementation (used data, ML algorithms) and the potential ethical risks, (b) to update the list of the ethical risks identified so far so that they meet technological advancement and pinpoint their interdependences, (c) to define scoring for ethical risks and to identify the timing in the system lifecycle when they are likely to occur. We consider that we have demonstrated the importance of ethics in the AER-based systems in online learning and we invite the readers to consider our practical guidelines in their future research and AER deployment.

Acknowledgement

This work was supported by a grant of the Petroleum-Gas University of Ploiesti, project number 11061/2023, within Internal Grant for Scientific Research.

References

ACM (Association for Computing Machinery). (2018). Code of ethics and professional conduct. https://www.acm.org/code-of-ethics, last accessed on 24 May 2023.

Alfoudari, A.M., Durugbo, C.M., Aldhmour, F.M. (2021). Understanding socio-technological challenges of smart classrooms using a systematic review, Computers & Education, Volume 173, 2021, 104282, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2021.104282.

Alwadei, A., Alnanih, R. (2022). Designing a Tool to Address the Depression of Children During Online Education, Procedia Computer Science, Volume 203, 173-180, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.07.024.

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. AI Magazine, 28(4), 15, https://doi.org/10.1609/aimag.v28i4.2065.

APA (American Psychological Association). (2017). Ethical principles of psychologists and code of conduct, https://www.apa.org/ethics/code, last accessed on 24 May 2023.

Arya, R., Singh, J., Kumar, A. (2021). A survey of multidisciplinary domains contributing to affective computing, Computer Science Review, Volume 40, 2021, 100399, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2021.100399.

Ashwin T.S., Guddeti R.M.R. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures, Future Generation Computer Systems, Volume 108, 334-348, ISSN 0167-739X, https://doi.org/10.1016/j.future.2020.02.075.

Azevedo, R., Bouchet, F., Harley, J. M., Feyzi-Behnagh, R., Trevors, G., Duffy, M., Taub, M., Pacampara, N., Agnew, L., & Griscom, S. (2011). MetaTutor: An intelligent multi-agent tutoring system designed to detect, track, model, and foster self-regulated learning. Proceedings of the

Fourth Workshop on Self-Regulated Learning in Educational Technologies. SRL&ET. Doi: https://doi.org/10.13140/RG.2.1.1334.6640.

Bakiner, O. (2022). What do academics say about artificial intelligence ethics? An overview of the scholarship. AI Ethics, https://doi.org/10.1007/s43681-022-00182-4.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest, 20, 1–68. doi:https://doi.org/10.1177/1529100619832930.

Barrett, L.F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. Social Cognitive and Affective Neuroscience, 12(1):1–23. https://doi.org /10.1093/scan/nsx060.

Bhardwaj, P., Gupta, P.K., Panwar, H., Siddiqui, M.K., Morales-Menendez, R., Bhaik, A. (2021) Application of Deep Learning on Student Engagement in e-learning environments, Computers & Electrical Engineering, Volume 93, 107277, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107277.

CDDO (Central Digital & Data Office). (2020). Guidance Data Framework, https://www.gov.uk/government/publications/data-ethics-framework, last accessed on 24 May 2023.

Cen, L., Wu, F., Yu, Z.L., Hu, F. (2016). A Real-Time Speech Emotion Recognition System and its Application in Online Learning, in Emotions, Technology, Design, and Learning, Editor(s): Sharon Y. Tettegah, Martin Gartmeier, vol. In Emotions and Technology, Academic Press, Pages 27-46, ISBN 9780128018569, https://doi.org/10.1016/B978-0-12-801856-9.00002-5.

Chen, C.M., Wu, C.H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance, Computers & Education, Volume 80, Pages 108-121, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2014.08.015.

COMEST (World Commission on the Ethics of Scientific Knowledge and Technology). (2019) Preliminary study on the ethics of artificial intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000367823, last accessed on 24 May 2023.

Cowie, R., (2015). Ethical Issues in Affective Computing', in Rafael Calvo and others (eds), The Oxford Handbook of Affective Computing, Oxford Library of Psychology, https://doi.org/10.1093/oxfordhb/9780199942237.013.006.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A.N., Raji, D., Rankin, J.L., Richardson, R., Schultz, J., West, S.M. and Whittaker, M. (2019). AI Now 2019 Report. New York: AI Now Institute.

Dai, C.P., Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review, Computers and Education: Artificial Intelligence, Vol. 3, 100087, https://doi.org/10.1016/j.caeai.2022.100087.

Dewan, M.A.A., Murshed, M. & Lin, F. (2019). Engagement detection in online learning: a review. Smart Learn. Environ. 6, 1, https://doi.org/10.1186/s40561-018-0080-z.

D'Mello S., & Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. 2, 4, Article 23, 39 pages. https://doi.org/10.1145/2395123.2395128.

D'Mello, S. K. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology, Journal of Educational Psychology, 105(4), 1082–1099, http://dx.doi.org/10.1037/a0032674.

D'Mello, S. K., & Graesser, A. C. (2015). Feeling, thinking, and computing with affect-aware learning technologies. In R. A. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), The Oxford handbook of affective computing, 419–434. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.032

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. Learning and Instruction, 29, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003.

Duo, S., Song, L.X. (2012). An E-learning System based on Affective Computing, Physics Procedia, Volume 24, Part C, 1893-1898.

EGE (European Commission, European Group on Ethics in Science and New Technologies). (2012). Ethics of information and communication technologies, Publications Office, 2012, https://data.europa.eu/doi/10.2796/13541.

Ekman, P. (1999). Basic emotions. In Handbook of Cognition and Emotion; Dalgleish, T., Power, M., Eds.; John Wiley&Sons Ltd.: Hoboken, NJ, USA.

Ekman, P.; Sorenson, E.R.; Friesen, W.V. (1969) Pan-cultural elements in facial displays of emotions. Science, 164, 86–88.

EUCFR (European Union Charter for Fundamental Rights). (2012). Charter of Fundamental Rights of the European Union , https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT, last accessed on 24 May 2023.

Ez-zaouia, M., Tabard, A., Lavoué, E. (2020). Emodash: A dashboard supporting retrospective awareness of emotions in online learning, International Journal of Human-Computer Studies, Volume 139, 102411, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2020.102411.

Faria, A.R., Almeida, A., Martins, C., Gonçalves, R., Martins, J., Branco, F. (2017). A global perspective on an emotional learning model proposal, Telematics and Informatics, Volume 34, Issue 6, Pages 824-837, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2016.08.007.

Feidakis, M. (2016). A Review of Emotion-Aware Systems for e-Learning in Virtual Environments, Editor(s): Santi Caballé, Robert Clarisó, In Intelligent Data-Centric Systems, Formative Assessment, Learning Data Analytics and Gamification, Academic Press, 217-242, ISBN 9780128036372, https://doi.org/10.1016/B978-0-12-803637-2.00011-7.

GDPR. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), https://gdpr-info.eu, last accessed on 24 May 2023.

Goldenfein, J. (2020). Facial Recognition is Only the Beginning. Public Books, Available at SSRN: https://ssrn.com/abstract=3546525, last accessed on 24 May 2023.

Goldie, P., Doring, S., Cowie, R. (2011). The ethical distinctiveness of emotion-oriented technology: Four long-term issues. In P. Petta, C. Pelachaud, R. Cowie (eds.), Emotion-oriented systems: The Humaine handbook. Berlin, Springer.

(Grawemeyer et al., 2016) Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes, Sergio Gutierrez-Santos, Michael Wiedmann, and Nikol Rummel. 2016. Affecting off-task behaviour: how affect-aware feedback can improve student learning. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16). Association for Computing Machinery, New York, NY, USA, 104–113. https://doi.org/10.1145/2883851.2883936

Guerini M., Stock O. (2005) Toward ethical persuasive agents in Proceedings of the IJCAI Workshop on Computational Models of Natural Argument, Edinburgh.

Harley, J.M., Bouchet, F., Hussain, M.S., Azevedo, R., Calvo, R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system, Computers in Human Behavior, Vol. 48, 615-625, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2015.02.013.

Hascher, T. & Edlinger, H. (2009) Positive Emotionen und Wohlbefinden in der Schule − ein Rubberlike über Forschungszugänge und Erkenntnisse [Positive emotions and well-being in school – an overview of methods and results], Psychologie in Erziehung und Unterricht, 56, 105-122.

Hascher, T. (2010). Learning and Emotion: Perspectives for Theory and Research. European Educational Research Journal, 9(1), 13–28. https://doi.org/10.2304/eerj.2010.9.1.13.

Hasnine, M. N., Bui, H.T.T., Tran, T.T.T., Nguyen, H.T., Akçapınar, G., Hiroshi Ueda, H. (2021). Students' emotion extraction and visualization for engagement detection in online learning, Procedia Computer Science, Volume 192, 2021, 3423-3431, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.09.115.

ICO (Information Commisioner's Office), Alan Turing Institute. (2020). Explaining decisions made with AI, https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/, last accessed on 24 May 2023.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE Global Initiative). (2018). Ethically Aligned Design – Version 2 for Public Discussion. https://standards.ieee.org/initiatives/autonomous-intelligence-systems/, last accessed on 24 May 2023.

IFOW (Institute for the Future of Work). (2020). Mind the gap: How to fill the equality and AI accountability gap in an automated world, https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force, last accessed on 24 May 2023.

Imani, M., Montazer, G.A. (2019). A survey of emotion recognition methods with emphasis on E-Learning environments, Journal of Network and Computer Applications, Volume 147, 102423, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2019.102423.

Iulamanova, A., Bogdanova, D., Kotelnikov, V. (2021). Decision Support in the Automated Compilation of Individual Training Module Based on the Emotional State of Students, IFAC-PapersOnLine, Volume 54, Issue 13, 85-90, ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2021.10.424.

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2.

Kazemitabar, M., Lajoie, S.P., Doleck, T. (2021). Analysis of emotion regulation using posture, voice, and attention: A qualitative case study, Computers and Education Open, Volume 2, 100030, ISSN 2666-5573, https://doi.org/10.1016/j.caeo.2021.100030.

Kulkarni, P. (2022). ML of Choosing: Architecting Intelligent Choice Framework. In: Choice Computing: Machine Learning and Systemic Economics for Choosing. Intelligent Systems Reference Library, vol 225. Springer, Singapore, https://doi.org/10.1007/978-981-19-4059-0_3.

Lavoué, É., Ju, Q., Hallifax, S., Serna, A. (2021). Analyzing the relationships between learners' motivation and observable engaged behaviors in a gamified learning environment, International Journal of Human-Computer Studies, Volume 154, 102670, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2021.102670.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529.

Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., Rincon, C. (2021). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. The Alan Turing Institute. https://doi.org/10.5281/zenodo.5981676.

Lin, F.R., Kao, C.M. (2018). Mental effort detection using EEG data in E-learning contexts, Computers & Education, Volume 122, 2018, 63-79, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2018.03.020.

Liu, S., Liu, S., Liu, Z., Peng, X., Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement, Computers & Education, Volume 181, 104461, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2022.104461.

Lyu, L., Zhang, Y., Chi, M.Y., Yang, F., Zhang, S.G., Liu, P., Lu, W.G. (2022). Spontaneous facial expression database of learners' academic emotions in online learning with hand occlusion, Computers & Electrical Engineering, Volume 97, 107667, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107667.

Mehrabian, A. (1996). Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament. Curr. Psychol, 14, 261–292.

Mohammad, S.M. (2022). Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis, Computational Linguistics, 48(2):239–278, https://doi.org/10.1162/coli_a_00433.

Moor, J.H. (2006). The Nature, Importance, and Difficulty of Machine Ethics, in IEEE Intelligent Systems, vol. 21, no. 4, 18-21, doi: 10.1109/MIS.2006.80.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, 29(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257.

Pekrun, R. (2005). Progress and open problems in educational emotion research. Learning and Instruction, 15(5), 497–506. https://doi.org/10.1016/j.learninstruc.2005.07.014..

Pekrun, R., Goetz, T., Titz W. & Perry R. P. (2002) Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research, Educational Psychologist, 37:2, 91-105, DOI: 10.1207/S15326985EP3702_4.

Picard, R.W. (1995). Affective Computing, M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, https://affect.media.mit.edu/pdfs/95.picard.pdf, last accessed on 24 May 2023.

Picard, R.W. (1997). Affective Computing. MIT Press, Cambridge, MA, USA.

Picard, R.W., Papert, S., Bender, W. et al. (2004). Affective Learning – A Manifesto, BT Technology Journal, 22: 253, https://doi.org/10.1023/B:BTTJ.0000047603.37042.33.

Qiao, X., Zheng, X., Sun, X., Li, S., Zhang, Y. (2022). Learners' States Monitoring Method Based on Face Recognition Technology, Procedia Computer Science, Volume 202, 172-177, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.04.024.

Rosenfeld, R. A. (1978). Anxiety and Learning. Teaching Sociology, vol. 5, no. 2, 1978, 151–66. JSTOR, https://doi.org/10.2307/1317061.

Ross, W.D. (1939). Foundations on ethics, Oxford, UK, Oxford University Press.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3):273–294, https://doi.org/10.1016 /0092-6566(77)90037-X.

Sadowski, J., Viljoen, S., Whittaker, M. (2021). Everyone should decide how their digital data are used — not just tech companies, Springer Nature, 595(7866):169-171. doi: 10.1038/d41586-021-01812-3. PMID: 34211184.

Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. J. Power (Eds.), Handbook of cognition and emotion (pp. 637–663). John Wiley & Sons Ltd. https://doi.org/10.1002/0470013494.ch30.

Sharma, K., Papavlasopoulou, S., Giannakos, M. (2022). Children's facial expressions during collaborative coding: Objective versus subjective performances, International Journal of Child-Computer Interaction, Vol. 34, 100536, https://doi.org/10.1016/j.ijcci.2022.100536.

Shelton, C. (2022). Complementary to Martin Hilbert course "Big Data, Artificial Intelligence, and Ethics", University of California, Coursera Plus.

Sikström, P., Valentini, C., Sivunen, A., Kärkkäinen, T. (2022). How pedagogical agents communicate with students: A two-phase systematic review, Computers & Education, Volume 188, 104564, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2022.104564.

Tanko, D., Dogan, S., Demir, F.B., Baygin, M., Sahin, S.E., Tuncer, T. (2022). Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23, Applied Acoustics, Volume 190, 108637, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2022.108637.

Um, E.R., Plass, J.L., Hayward, E.O., and Homer, B.D. (2012). Emotional design in multimedia learning, Journal of Educational Psychology, 104(2), 485–498. https://doi.org/10.1037/a0026609.

UNESCO. (2021). Recommendation on the Ethics of AI, https://unesdoc.unesco.org/ark:/48223/pf0000380455, last accessed on 24 May 2023.

Utami, P., Hartanto, R., Soesanti, I. (2019). A Study on Facial Expression Recognition in Assessing Teaching Skills: Datasets and Methods, Procedia Computer Science, Volume 161, 544-552, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.154.

Vidanaralage, A.J., Dharmaratne, A.T., Haque, S. (2022). AI-based multidisciplinary framework to assess the impact of gamified video-based learning through schema and emotion analysis, Computers and Education: Artificial Intelligence, Volume 3, 100109, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2022.100109.

Xu, T., Zhou, Y., Wang, Z., Peng, Y., (2018). Learning Emotions EEG-based Recognition and Brain Activity: A Survey Study on BCI for Intelligent Tutoring System, Procedia Computer Science, Volume 130, Pages 376-382, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.04.056.

Yadegaridehkordi E., Noor N.F.B.M., Ayub M.N.B., Affal H.B. & Hussin N.B. (2019). Affective computing in education: A systematic review and future research, Computers & Education, doi: https://doi.org/10.1016/j.compedu.2019.103649.

Yang, D., Alsadoon, A., Prasad, P.W.C., Singh, A.K., Elchouemi, A. (2018). An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment, Procedia Computer Science, Volume 125, 2-10, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.12.003.

Zeidner, M. (1998). Test anxiety: The state of the art. Plenum Press.

International Journal of Computational Intelligence Systems (2024) 17:230
https://doi.org/10.1007/s44196-024-00638-x

RESEARCH ARTICLE

Towards Integrating Automatic Emotion Recognition in Education: A Deep Learning Model Based on 5 EEG Channels

Gabriela Moise[1] · Elia Georgiana Dragomir[1] · Daniela Șchiopu[1] · Lidia Angelica Iancu[1]

Received: 13 May 2024 / Accepted: 19 August 2024

Abstract
In a technologically advanced world, artificial intelligence has impacted all fields of activity. The augmentation of online learning by means of emotion recognition systems raises new challenges in terms of obtaining high-performance systems and in interpreting the results. The paper aims to investigate the usage of automated emotion recognition in learning and to develop a deep learning model based on physiological data to recognize emotions often encountered in classrooms. So, an 1D-CNN model based on physiological data is used to recognize seven emotions: boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety. These emotions are described according to the PAD model and the 5 EEG signals, FP1, AF3, F7, T7, FP2, are taken from the DEAP dataset to train and to test the convolutional neural network model. The high accuracy we obtained (i.e. boredom—99.64%, confusion—99.70%, frustration—99.66%, curiosity—99.80%, excitement—99.91%, concentration—99.70%, anxiety—99.21%) proves that the use of signals obtained via only five channels is sufficient to recognize the presence of emotions. Furthermore, an improved method of analysis based on LIME is proposed and used to obtain reliable explanations for the predictions of our model.

Keywords Emotion recognition · Affective learning · Deep learning · Biophysical data

1 Introduction

Artificial intelligence has become ubiquitous in today's world, and, therefore, its applications can be found in a wide range of domains, education included. With the adoption of artificial intelligence (AI) in education, there have emerged both new opportunities to improve the teaching and learning processes, and challenges, of which we mention ethical concerns, information security, data manipulation, etc.

There have been identified four roles of AI in education (AIEd): as an "*intelligent tutor*", where one can include most of the applications, such as the intelligent tutoring systems (ITS) or the adaptive learning systems; as an "*intelligent tutee*", given the fact that AI facilitates students to be tutors; as an "*intelligent learning tool/partner*" making it easier for the students to focus on high-level tasks while helping them solve low-level tasks; and as a "*policy-making advisor*" providing tools that help policymakers to analyse and understand the problems in education, as well as to find solutions [1]. In a position paper, the authors show the paradigm changes that AIEd supported: "*AI-Directed*", in which one can see "*learner-as-recipient*"; "*AI-Supported*' with "*learner-as-collaborator*", and "*AI-Empowered*" with "*learner-as-leader*" [2]. In the "*AI-directed*" paradigm, the machine guides the learning process, and the learner has the role of an AI-services recipient. The "*AI-supported*" paradigm gives the student the role of collaborator of the machine and, as a result, the system is a supporting tool. Whereas in the third paradigm, "*AI-Empowered'*, AI augments human intelligence and the machine interacts with human intelligence synergistically in order to provide personalised learning. The third paradigm represents the development trend of AI applications in education, aiming "to empower learners to take full agency of learning" [2]. Undoubtedly, in the whole context of the integration of AI in education, the complexity of the learning and teaching processes must be considered, in which the emotions of the participants in the process play a major role.

Pekrun calls the classroom "*a space of emotions*" and argues both the influence of the emotional state of the

✉ Gabriela Moise
gmoise@upg-ploiesti.ro

[1] Department of Computer Science, Information Technology, Mathematics and Physics (ITIMF), Petroleum-Gas University of Ploiești, 100680 Ploiești, Romania

students on the process and, conversely, the change of the state due to the process [3]. The reciprocal relation between students' emotions and learning processes, academic achievements, students' well-being, their motivation and engagement has been proven in many research studies [4, 5]. Moreover, the teachers' emotions affect those of their students, the teaching process, the teacher–student relationship, their students' cognition, motivation and the outcome, practically the whole classroom [6, 7].

In technology-mediated learning, machines take over some of the teachers' tasks along with their ability to have and express emotions, on the one hand, and to recognize and manage emotions, on the other. In 2004, a new field of study appeared, one related to AI, emotions, and the learning process. "*Affective learning*" brings together multiple perspectives regarding students' emotions, machines capable of feeling, recognizing, and expressing emotions, as well as learning systems which incorporate emotions [8]. Automated emotion recognition (AER) has been integrated in learning systems and as a result of their increasing use, new issues have arisen, which demand both maximum attention from the scientific communities and sustainable solutions. In [9], there are investigated both the benefits of using automated emotion recognition in online learning and the ethical risks that may arise as a result of their use. With respect to the benefits of AER in online learning, the authors highlight six categories of AER roles in education, as well as possible benefits for each of these categories: ITS as pedagogical agents with emotions (which allows customized feedback, effectiveness of learning, support for teachers); engagement and motivation support for learners and teachers (awareness of emotional states); learning and teaching processes assessment (fraud-free results, teachers' awareness of the role of emotions in teaching); emotional favourable learning environment; support for students with special needs. As far as possible ethical risks are concerned, there have been identified 16 risk classes, among which one can mention poor results of machine learning (ML) models, non-transparent and unexplainable models' outcomes, possible bias and discrimination of predictions, dependence on the machine, etc. In the recently adopted regulatory document, the AI Act, the main concerns of AER systems usage are highlighted, namely, limited reliability and generalizability, the possibility to obtain discriminatory results, and to be intrusive in the life of the persons [10].

Although machine learning models make accurate predictions or provide a solid basis for the decision-making process, there are still uncertainties about how the results of the proposed solutions are obtained. Machine learning models must become tangible for users so that they may increase the efficiency of ML applications and the confidence of their usage in various domains. ML models should not only be selected based on performance criteria, but more partially quantifiable important criteria also need to be considered when building ML-based systems as Dosh-Velez and Kim highlighted in [11]: "*safety*, *non-discrimination*, *the right to explanation*, *avoiding technical debt*, and *interpretability*". Leveraging the benefits and risks of the usage AER applications in education, we claim the need to develop high-accuracy emotion recognition systems having at least the following characteristics: reliable, non-discriminatory, and less intrusive. Also, we state the necessity to provide explanations for the predictions generated by ML models.

1.1 Scope

The paper aims to investigate the usage of automated emotion recognition in learning, to develop a reliable and performant AER model and to provide explanations for the predicted results.

So, we perform a brief examination of the automated recognition of emotions as encountered and manifested within the learning process. Also, we develop a reliable and performant AER model and provide explanations for the predicted results.

There are many ways to detect emotions: facial expressions, the tone of voice, body language, biophysical data (EEG—electroencephalogram, GSR—galvanic skin response, heart rate, respiration rate), etc. [12–16]. One of the most popular and reliable techniques for AER is using EEG signals [16].

Based on the public benchmark Dataset for Emotion Analysis using Physiological signals (henceforth DEAP), a dataset provided by Koelstra et al. [17], we experiment a convolutional neural network (CNN) emotions recognition model designed by Akter et al. [18] based on biophysical data related to the following seven emotions: *boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety*. We select the 1D-CNN model from [18] because it is the best performing EEG-based model for AER found in the literature as we show in Sect. 3.5. The model is trained on four datasets: one uses 14 EEG (electroencephalogram) channels, as in [18] to validate the performance of the model, two datasets use 10 EEG channels each, as in [19], and the last one makes use of 5 EEG channels derived from [19]. Our goals are to decrease the complexity of input data through reducing the number of used EEG channels and to preserve the high accuracy of the machine model. Furthermore, we provide explanations for the obtained results so that to address the non-transparency of ML models. The explanations are carried out using an analysis method based on LIME (Local Interpretable Model-agnostic Explanations).

1.2 Findings

We prove that 5 EEG channels are sufficiently for the CNN model to be highly accurate (i.e. boredom—99.64%, confusion—99.69%, frustration—99.65%, curiosity—99.79%, excitement—99.90%, concentration—99.69%, anxiety—99.20%). The key shortcoming of performant AER models using EEG signals found in the literature is the usage of data extracted from many EEG channels leading to many attributes in the data set [18, 44, 45]. So, in these situations there are recorded high consumption of resources and an increasing training time.

Comparing the outcomes obtained by our approach with the results of the same model (1D-CNN) trained on data extracted from two sets of 10 EEG channels we observe a negligible depreciation of the accuracy in some cases, meanwhile the training time decreases considerably, except one case in which there is a slight increase in training time. Another advantage of using only 5 EEG channels is the possibility of building a low-cost wearable device for a real-time recognition of emotions.

The LIME-based analyse method applied to predictions' explanations is relevant to the domain, as the issue of explanations of AER results has been previously addressed, either vaguely or tangentially in the specialised literature; most studies focus on obtaining high-performance machine learning (ML) models and less on the explainability.

1.3 Outline

This paper is structured as follows: an overview of AER use in education, followed by materials and methods we used in our experiments (emotions models, the PAD description for the seven emotions encountered in the learning process, EEG signals, DEAP dataset, a short review of most performant AER based on DEAP dataset and about explicability). Our approach of AER and the model explicability is detailed and validated by a series of experiments. The final section is dedicated to conclusions that pinpoint both the strengths and limitations of our study, our achievements and future research ideas.

2 AER in the Educational Process

Indisputably, there are inextricable relationships between emotions and learning. In the educational process, one must consider both students' emotions and those of the teachers. Human beings can recognize emotions and act accordingly. Within online processes, machine-based emotions recognition tools used to improve learning are necessary, but their usage in education represents a controversial subject because of the involved risks [9]. This paper does not aim to analyze the benefits and risks of AER usage in education, but to make a statement that AER must be used with caution, transparently and with respect for ethics.

In what follows, we carry out a brief examination of the possible roles of AER in the educational process.

One of the most famous projects related to the computerized monitoring of students' affective state and providing appropriate feedback is "Learning Companion" (2000–2004) developed by Affective Computing Group, MIT Media Lab (https://www.media.mit.edu/projects/learning-companion/overview/). Kort, Reilly and Picard proposed in the project a model able to reflect the dynamics of emotions encountered in the learning process [20]. The goal of their model was to develop a "*computerized Learning Companion*" capable of recognizing the affective state of the learners and, consequently, of reacting appropriately. The emotions space was divided into four quadrants, each one reflecting a set of emotions and a phase of the learning process. In the learning process, a student experiences several emotional states associated with various stages of learning, for example: I. happiness—they are engaged in discovery learning; II. confusion—misunderstanding occurs; III. frustration, anger—awareness of error; IV. hopefulness—comeback to construct the understanding of the topics. For each phase, the learner needs adequate emotional support that, in the past, used to be provided by a teacher or their peers.

A prototype of an e-Learning model which includes the tracking of the affective states of the learners is described in [21]. The model considers the learning goals, the contextual information and cognitive abilities of the learners and provides personalized feedback according to the affective state of the learners. To validate the model, the authors conducted an experimental study for two weeks on a single subject, an undergraduate student from the computer science department. The learning content was recommended to the student both considering his emotional state and ignoring it. The results indicate that emotion-aware content recommendation requires fewer human interventions in the learning process, 11 interventions in the case of emotion-aware recommendation compared to 21 in the case of non-emotion-aware recommendation [21].

An intelligent tutoring system, called Affective AutoTutor, uses the affective and cognitive states of the students and responds accordingly. The system presented in [22] comprises two affect-detection augmented versions of AutoTutor, an intelligent system that helps students according to their cognitive abilities. The results obtained with these tutors show that students with more knowledge background do not need feedback according to their emotional states. On the other hand, for students with less knowledge, emotional support is important.

A students' engagement recognition model is developed in [23]. Based on facial expressions, the model performs a

classification in two categories: engaged and disengaged. The model can be integrated in the learning systems to facilitate the learning process. In another study, the emotional reactions of the learners are analysed and there is proposed an intelligent agent (ERPA) capable of predicting the emotional reaction of students after getting their grades in an exam [24]. ERPA predicts correctly the emotional reactions in 28 cases from 34.

The learners' engagement is a key factor in the efficiency of the learning process. In [25], the authors propose an emotional engagement detector using a hierarchical semi-supervised model. Three states are detected: satisfied, bored, and confused. The detector was tested both in instructional and assessment settings. Kadar et al. highlight in [26] the possibility of use AER in preventing students' dropout. The presented scenario includes four case studies performed in the classroom: the analysis of students' gait and posture when they enter or leave the classroom, the eye tracking and the facial emotion detection during the lessons (the students are sitting), and emotional states recording. All these cases are integrated to manage the affective states of learners to mitigate school dropout.

The case of e-learning systems and the role of cognitive emotions in the online learning process are approached in [27]. The emotions are detected from facial expressions recorded during a video-lecture and a conversation with the teacher. The obtained results confirm that cognitive emotions have a key role in a successful deployment of an online learning process.

A detailed exploration of AER roles in education is presented in [9]: as Intelligent Tutoring Systems with emotional capabilities; as support for engagement and motivation both for students and teachers, for assessment both for learning and the teaching process, for students with special needs, for building comfortable classroom environments. We must mention a notable study in [28] the authors performed a systematic literature review to debate both the educational and technical aspects of AER in online learning. 117 studies from the period 2010 to 2024 were analysed and the following findings were found: "(1) most articles were proposed to design and test systems for AER, neglecting how to apply such affective recognition systems in specific online learning environments. (2) Most of the research merely focused on the output of accomplishing emotion recognition tasks without considering the differences in online education contexts, including pedagogies, disciplines and academic levels. (3) The majority of studies on AER have focused more on discrete emotion models, primarily using basic emotions and emotion polarities, rather than learning-related emotions and the intensity of emotions; public databases were used more frequently with facial expression being the most used emotion recognition channel, contrasting with self-built databases where textual sentiment analysis prevailed. Physiological signals and multimodal fusion are rare used in both types of databases. (4) While many algorithm models exhibited impressive accuracy in AER, their interpretability and practical applicability in real learning scenarios were notably hindered by the limitations of the databases employed" [28].

In Table 1, there are resumed the examples of AER usage in education.

Table 1 AER usage in the educational process

A detailed exploration of roles of AER in educational process	[9]
A "*computerized Learning Companion*" capable to react appropriately at the affective state of learners. The model was developed in "Learning Companion" project, Affective Computing Group, MIT Media Lab	[20]
Experimental study carried out in a Shanghai online college: learners' emotions are recognized based on physiological signals. The results show the positive impact of AER on the learning process	[21]
Controlled experiments performed with Affective AutoTutor: the affect recognition is based on multimodal data (facial, text, body movements). Usage of Affective AutoTutor proves "dramatic improvements in the learning compared to the original AutoTutor system"	[22]
Deep learning model used to maintain the engagement of learners during learning sessions via technology. The engagement is recognized based on facial expressions	[23]
ERPA (Emotional Response Predictor Agent), an intelligent agent able to predict learners' reactions in the online learning environment. ERPA uses personal attributes of learners as personality (extraversion, lie scale, neuroticism, psychoticism) and sex (male, female)	[24]
Controlled experiments in the context of online math learning used to detect students' engagement based on their appearance expressions and context—performance data (e.g. user profile, data extracted from online learning platform)	[25]
A virtual scenario in a smart classroom used to detect the emotional states of students. There are proposed four cases to be analysed based on gait and posture; eye gaze; facial expression; data integration. The goal is to prevent school dropout	[26]
Experiments testing the reliability of software to detect and classify students' emotions based on facial expressions recorded in two situations: video lecture and conversation with teacher. Cognitive emotions monitoring can provide information related to the quality of the learning process	[27]
A comprehensive review of AER to include both the technical and pedagogical aspects of AER in online education	[28]

3 Materials and Methods

3.1 Emotions Models

The term "emotion" does not have a universally agreed-upon definition in the specialised literature. Fehr and Russell [29], Plutchik [30] and many others highlighted the challenge in giving an exhaustive definition of the concept. The effort to define *emotion*, formerly called *passion*, or *pathos*, dates back to ancient times. Philosophers and psychologists have categorized it as behaviour, mental event or type of judgment. Nevertheless, most of the definitions often fell short of covering all aspects, and therefore, in the last decades, there have been efforts to create a more inclusive definition of emotion (or affect) that may incorporate behaviour, physiological, and mental events. Izard defines emotion as a *complex process* involving neurophysiological, motor-expressive, and phenomenological elements [31].

There are three models widely used to represent emotions: the *discrete* model, the *dimensional* model and the *componential* model.

In the paper entitled *Pan-cultural elements in facial displays of emotions*, Ekman, Sorenson and Friesen highlight six basic emotions: "*happiness, fear, disgust-contempt, anger, surprise, sadness*" [32]. Few years later, Izzard identifies and defines 10 fundamental emotions, as follows: "*interest-excitement, joy, surprise, distress, anger, disgust, contempt, fear, shame, and guilt*" [33]. In 1999, Ekman described the characteristics based on which one can distinguish basic emotions and considered the following 15 basic emotions: "*amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure and shame*" [34]. Basic emotions may explain compound emotions.

The dimensional model describes an emotion using more dimensions, defined in either a two-dimensional space (*Valence* and *Arousal*—VA) or in a three-dimensional space (*Pleasure/Valence*, *Arousal*, and *Dominance*—PAD). *Valence* dimension indicates the positivity or negativity of an emotion, ranging from unpleasant feelings to pleasant ones. The *arousal* dimension reflects the level of excitement conveyed by the emotion, ranging from low/sleepness to intense excitement. The *dominance* dimension expresses the degree of control or influence associated with the emotion.

One of the best-known dimensional models is the "*circumplex model of affect*" proposed by Russell in 1980 [35]. A specific affect state is expressed as a point in two-dimensional space, valence (pleasure) and arousal (activation). 28 affect states are represented along a circle starting with *happy* at 7.8° and ending with *pleased* at 353.2°. The 28 states identified and analysed by Russell are: "*happy, delighted, excited, astonished, aroused, tense, alarmed, angry, afraid, annoyed, distressed, frustrated, miserable, sad, gloomy, depressed, bored, droopy, tired, sleepy, calm, relaxed, satisfied, at ease, content, serene, glad, and pleased*" [35].

Mehrabian and Russell proposed in 1977 a PAD model able to distinguish between *anger* and *fear* using the dominance dimension [36–38] and there was stated that anger is characterized by positive dominance, whereas fear is defined by negative dominance.

In the componential model proposed by Plutchik, a complex emotion is defined as a combination of basic emotions [30]. The spectrum of emotions is visualised using a wheel of emotions, and comprises eight basic emotions (i.e. *anger, fear, sadness, disgust, surprise, anticipation, trust, and joy*), as well as a scale of emotion intensity. Complex emotions are created using combinations of the eight above-mentioned core emotions.

3.2 The PAD Description for the Seven Emotions Examined in Our Work

In our study, we consider the following seven emotions more often encountered in the learning process: *boredom*, *confusion*, *frustration*, *curiosity*, *excitement, concentration*, *anxiety* and we use a PAD model to express each emotion. Moreover, we take into account the results presented by Russel & Mehrabian in the paper entitled *Evidence of a Three-Factor Theory of Emotions*, in which emotions are described in terms of: pleasure–displeasure, arousal–non-arousal, and dominance–submissiveness. In Table 2, there are presented mean and standard deviation values on the [−1, 1] range for pleasure, arousal, and dominance dimensions provided by Russel & Mehrabian for the seven emotions: *boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety* [36].

The mean values for pleasure, arousal, and dominance of the seven emotions are graphically represented in Fig. 1.

3.3 EEG Signals

Automatic emotion recognition (AER) systems use data extracted from both non-physiological and physiological signals. Non-physiological signals refer to the external observations such as those related to the tone of voice/speech, gestures, facial expressions, text, etc. These signals are not reliable for the AER systems and cannot provide a high level of recognition accuracy, because factors such as an individual's age, culture, sex, and habit may influence

Table 2 Mean and standard deviation PAD values for seven emotions [36]

Emotion	Pleasure		Arousal		Dominance	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Boredom	−0.65	0.19	−0.62	0.24	−0.33	0.21
Confusion	−0.53	0.2	0.27	0.29	−0.32	0.28
Frustration	−0.64	0.18	0.52	0.37	−0.35	0.3
Curiosity	0.22	0.3	0.62	0.2	−0.01	0.34
Excitement	0.62	0.25	0.75	0.2	0.38	0.29
Concentration	0.42	0.25	0.28	0.27	0.39	0.31
Anxiety	0.01	0.45	0.59	0.31	−0.15	0.31

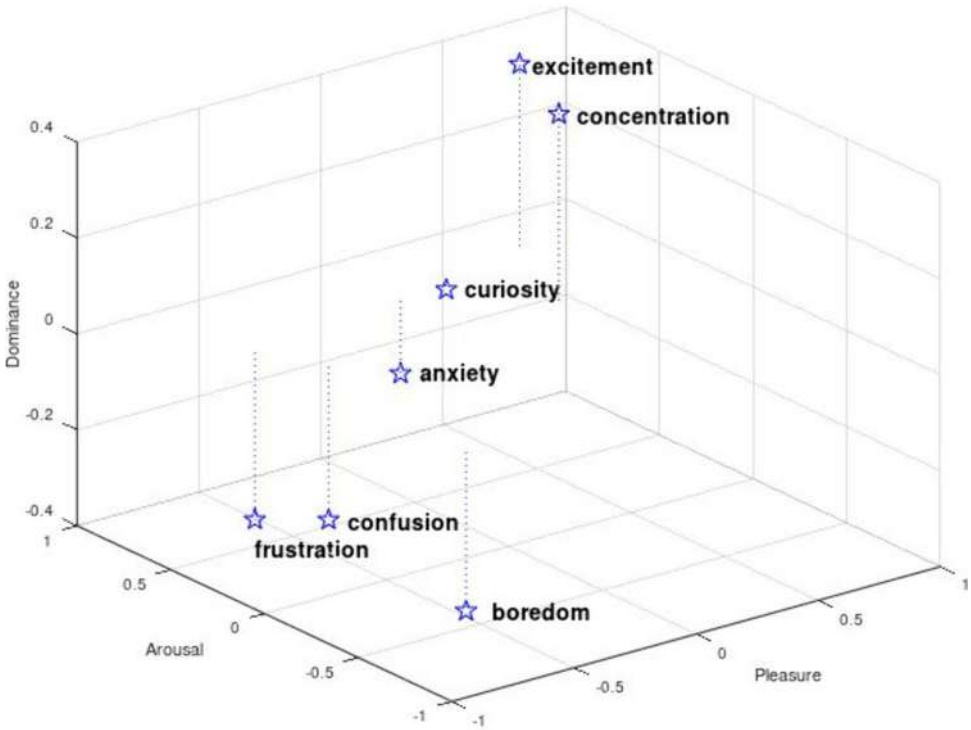

Fig. 1 PAD representations for the seven emotions (according to mean values, as provided in [36])

emotion recognition [39, 40]. Moreover, an individual can conceal their true emotional state, which means that a non-physiological-based AER can be deceiving.

Conversely, physiological feature, such as electroencephalography (EEG), skin temperature, skin conductance (SC), respiration rate, blood volume pulse, electromyography (EMG), eye gaze are more reliable and, therefore, superior to non-physiological attributes when it comes to automatic emotion recognition [40, 41] It has been shown that there is a close relationship between an individual's emotional state and their brain activity and, consequently, EEG signals are increasingly used in the development of the AER systems [39, 41, 42]. The range of EEG frequency measured on the scalp varies between 1 and 100 Hz. EEG is a composite signal with five frequency sub-bands relevant to an individual's mental state: delta waves (< 4 Hz), theta waves (4–8 Hz), alpha waves (8–13 Hz), beta waves (13–30 Hz), and gamma waves (> 30 Hz). The ends of the intervals vary by 1–2 Hz in different studies. According to [40, 43] emotions are strongly connected to gamma and beta sub-bands and mediums connected to alpha sub-bands. On the other hand, emotions are low related to theta sub-band. Although EEG-based AER systems are highly accurate, Rahman et al. reveal the limits of these systems due to the variability of an individual's emotional states and to various factors which may influence EEG signals, such as time of the day of EEG signals acquisition [41].

EEG signals are acquired using a standard electrode system e.g., 10/20 system presented in (https://trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf). The positions of electrodes placements in 10–20 international system are shown in Fig. 2.

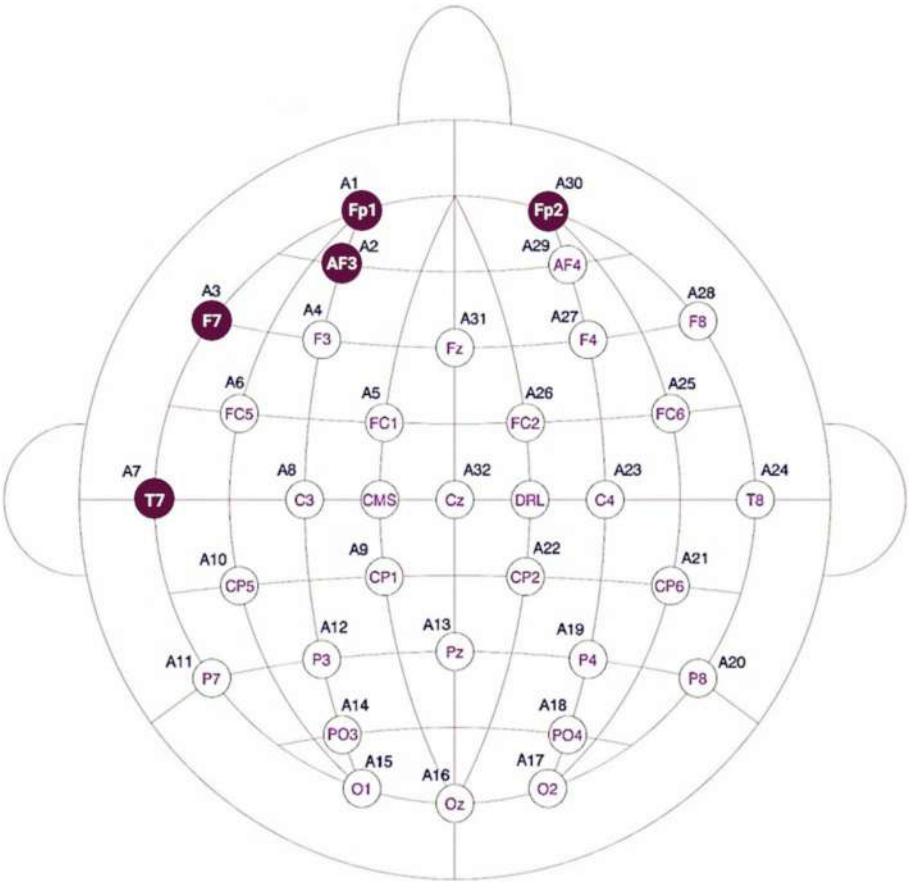

Fig. 2 Electrode placement in DEAP, 10/20 international system

The position of each electrode is specified by means of a letter and a number: the letter specifies the lobe (F—frontal, T—temporal, C—central, P—parietal, O—occipital) and the number—the position in each hemisphere of the brain (2, 4, 6, 8—right hemisphere; 1, 3, 5, 7—left hemisphere). The letter 'z' specifies the position on the midline. 10/20 refers to the distance between electrodes: 10% or 20% of the front–back or left–right distance of the skull.

3.4 DEAP Data Set

DEAP dataset is a multimodal dataset containing electroencephalography (EEG) signals and peripheral physiological signals (galvanic skin response—GSR, respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram—EOG), acquired from 32 subjects while they were watching 40 music videos [17]. Each participant rated valence, arousal, dominance, like/dislike with a float value from 1 to 9 and familiarity with an integer value from 1 to 5. EEG signals (at a sampling rate of 512 Hz) were recorded by using AgCl 32 electrodes positioned according to the international 10/20 system and 12 peripheral signals were acquired by using sensors on the face, neck and the left hand, as well as a respiration belt.

In our paper, we use preprocessed data, a selection from 32 EEG channels from [17] downsampled to 128 Hz, with artefacts removed, to which it was applied a bandpass frequency filter from 4.0 to 45.0 Hz and averaged to a common reference.

3.5 Emotion Recognition Models Based on DEAP Dataset

DEAP dataset is an intensely used database in AER systems (2865 citations of paper [17] as of August 2, 2024 reported by IEEE engine). Our aim is to obtain high accuracy models for the recognition of the seven chosen emotions, even we use complex ML models, which are not intrinsically interpretable.

To achieve it, we have conducted research on ML models for emotion recognition that use EEG values from DEAP database and selected only the models with accuracy closed to or greater than 90%.

In [44], the authors use the raw EEG data and an 1D-CNN with Conv-BN layer fusion quantization technique to classify valence, arousal and valence-arousal. The obtained model is hardware-friendly and the accuracy values varies between 93.16% and 96.62%. Another approach is presented in [45], EEG signals are preprocessed, only the Gamma

band was considered, after that, partial mutual information (PMI) was used for feature extraction. Two connectivity feature maps (CFM) were built, one in 2D and another in 3D forms. Binary classifications were performed for valence and arousal using 2D-CNN for 2D CFM, respectively, 3D-CNN for 3D CFM. Accuracy values are around 91–92%. A 3D feature structure (in terms of frequency, time and spatial domains) containing EEG signals information was used in [46]. A new CNN was designed, multiscale frequency–time–spatial convolutional model—MSFTSCNN, to classify valence and arousal. The method was tested on two datasets DEAP and MOHNOB-HCI. The obtained accuracies for DEAP is 93.82% for arousal, respectively, 94.48% for valence in the case of DEAP. High values for accuracies were obtained in [47] and [49], about 98%, for valence and arousal, and in addition in [49] and for dominance. In both studies there used data from all EEG channels. The ML models used are a GJFusion network in [47] and residual Long-Short Term Memory (ResLSTM) in [49]. In [48] there used raw EEG without data preprocessing and feature extraction and an end-to-end framework, Spatiotemporal Symmetric Transformer Model (STS-Transformer), for emotion recognition.

The most performant strategy for emotion recognition is presented in [18]. Using only 14 EEG channels, Fast Fourier Transformation for feature extraction and two CNN models (one heavily parametrized and the other lightly parametrized), the authors obtain high values for accuracies, between 97.80% and 99.89%.

Table 3 shows the best performing models identified after having examined the models considered to be the most relevant ones.

The best performance for AER models was obtained by Akter using a 1D-CNN model with four convolutional layers and three dense layers [18]. As a result, we considered it when making our predictions with respect to the presence/absence of the seven emotions, i.e. *boredom*, *confusion*, *frustration*, *curiosity*, *excitement, concentration,* and *anxiety* in the learning process.

Comparing with the aforementioned methods, our approach uses only five channels, recognizing seven emotions described in PAD dimensions, and in addition providing explanations for the predictions. Reviewer 1, pct. 2

Table 3 AER models with high performance

References	Classifiers	Performance (%)
[44]	1D-CNN + Conv-BN layer fusion quantization technique Raw EEG data	Accuracy Valence—96.62 Arousal—98.18 Valence-Arousal—93.16
[45]	2D-CNN Partial Mutual Information	Accuracy Valence—91.35 Arousal—92.18
	3D-CNN Partial Mutual Information	Accuracy Valence—91.71 Arousal—91.99
[46]	MS-TSCNN (multi-scale one-dimensional convolutional model) Differential Entropy	Accuracy Arousal—93.82 Valence—94.48
[47]	GJFusion (Graph Joint Fusion) network Raw EEG	Accuracy Valence—98.24 Arousal—98.38
[48]	STS-Transformer (Spatiotemporal Symmetric Transformer Model) Raw EEG	Accuracy Valence—89.96 Arousal—86.83
[49]	Improved capsule network and residual Long-Short Term Memory (ICaps-ResLSTM) Spatial feature extraction	Accuracy Arousal—98.06 Valence—97.94 Dominance—98.15
[18]	1D-CNN (4 Conv + 3 Dense) FFT (Fast Fourier Transform) FE-3 (Feature extension three different scores)	Accuracy Valence—99.89 Arousal—99.83
	1D-CNN (2 Conv + 1 Dense) FFT (Fast Fourier Transform) FE-3 (Feature extension three different scores)	Accuracy Valence—99.22 Arousal—97.80

3.6 Interpretability vs. Explainability

Interpretability and *explainability* are two concepts often related to the *understanding* of machine-learning models by humans, be them experts, or non-experts. The researchers still debate the definitions of interpretability and explainability; some consider them to represent the same thing, whereas others make a clear distinction.

The ability to explain decisions in artificial intelligence does not represent a new topic. In 1977, Scott et al. proposed a Production-base Consultation System augmented with Explanation Capabilities [50]. They argued for the expansion of systems with explanatory capabilities with the need for users to have access to as much knowledge of the system as possible. Users should be able to receive comprehensive answers to questions such as how a decision was made, what information was used, why the system failed, etc. [50].

Interpretability has become a hot topic in recent years. Ribeiro et al. give a definition for "*explaining a prediction*" by means of visual or textual artifacts that generate a "*qualitative understanding*" of the relationship between the attributes of an instance (the input) and its related prediction (the output) [51]. A strong remark made by authors refers to the user's background and the experience necessary to consider when one approaches the interpretability in ML models.

Murdoch et al. provide the following definition of interpretable ML, also referred to as explainable ML, transparent ML, or intelligible ML: "*the extraction of relevant knowledge from a machine learning model concerning relationships either contained in data or learned by the model*" [52]. The knowledge must be presented in formats that are tangible for various audiences: graphics (visualisation), natural language, or mathematical equations. Guegan makes a clear difference between interpretability and explicability in ML, as follows: interpretation deals with understanding of prediction, *why* this output was obtained, whereas explicability refers to *how* the ML model works [53].

The *global interpretability* consists in understanding how the whole ML model works, whereas *local interpretability* refers to understanding the reasoning of a certain decision.

Closely related, explainability or interpretability is hard to achieve in case of complex ML models. The more complicated, the more difficult is the task to provide a justification of their results. On the other hand, high interpretable ML models, such as linear regression, decision trees or SVMs provide low performances and, as a result, researchers use various approaches for non-interpretable models so that to make their results understandable to people.

There are *model-specific* and *model-agnostic methods*. Model-specific methods do not work for any model. For example, in linear regression, weights are used for interpretability, and the contribution of a numerical feature to an individual prediction is given by the weight—feature's product value for that instance. Model-agnostic methods are usually applied after the training of the model and they work for any machine-learning model. A comprehensive study on methods and metrics used in the field of machine-learning interpretability can be found in [54].

Given the fact that we have chosen a non-intrinsically interpretable model for emotion recognition (CNN), we considered appropriate to use a predictions' analysis method based on LIME.

LIME (Local Interpretable Model-agnostic Explanations) is an algorithm proposed in [51] that explains individual predictions through training of the local surrogate models. LIME is a feature-based method and provide for each feature a value (a score) representing the importance of it (feature) in the prediction.

The idea of LIME is quite simple, and it consists of the following steps:

- Select an instance (x, y), where $x=(x_1, x_2, \ldots, x_p)$ is the vector of feature values and y is the prediction for x.
- Build an input data set $\{x^1, x^2, \ldots, x^k\}$ in the neighbourhood of the instance x, perturbing the feature values. For each perturbed x^i obtain the prediction y^i. So, a new data set is generated (x^1, y^1), (x^2, y^2), $\ldots(x^k, y^k)$.
- Each x^i receives a weight according to its proximity to the instance.
- Using the new data set (x^i, y^i), $i=1,k$ an intrinsic interpretable model (a surrogate model as linear regression, Lasso, decision tree, etc.) is trained.
- The interpretable model approximates the behavior of the non-interpretable model in the proximity of the instance x, but not global. This new interpretable model provides an explanation for the instance x and its prediction y.

LIME generates a value for each feature and has the advantage of generating human-friendly explanations. This is why we choose LIME. Conversely, its instability is a possible weakness: the width of the neighbourhood influences the results, and, therefore, different interpretations are obtained with each run of the algorithm. In order to solve this problem, we propose an analysis method based on LIME to provide explanations for the results of our AER.

4 Our Approach for AER to Be Used in Learning Process

The overview diagram of emotion recognition using EEG signals and results' explanation considered in our experiments is presented in Fig. 3. The first stage in our approach consists of selecting the most relevant EEG channels for emotion recognition, after that we performed the features' extraction. In next stage, we adapted the most performant

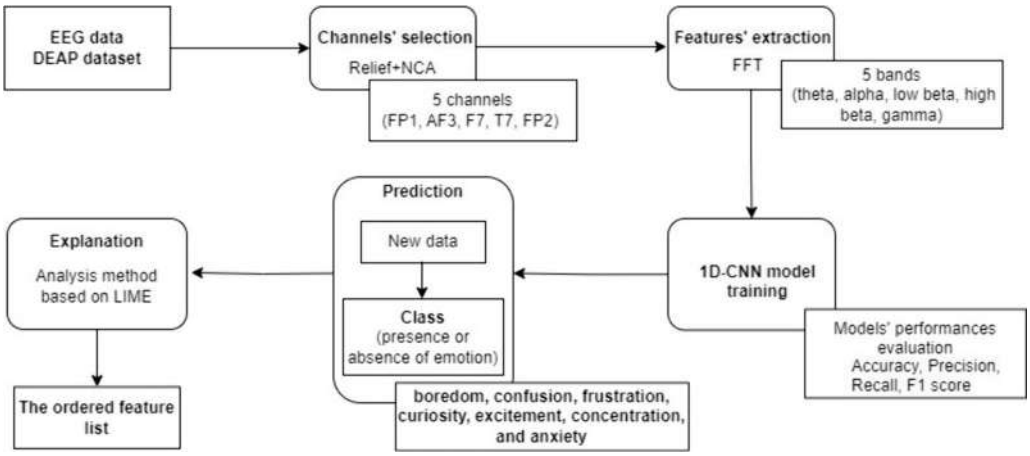

Fig. 3 Our framework for AER

Table 4 14 EEG channels used in [18]

FP1	FP2
AF3	Fz
F3	F4
F7	F8
FC1	C4
P3	P4
PO3	PO4

model for emotion recognition found in [18] for our data set and made predictions for seven emotions: boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety. Finally, we use a LIME-based method to generate explanations for the results predicted by the model.

4.1 EEG Channels Selection

The AER model designed by Akter et al. in 2022 has the highest performance, and uses 14 EEG channels from DEAP dataset, as presented in Table 4 [18].

Our goal was to provide explications for the predicted labels, and, as a result, we conducted an investigation to reduce the complexity of the input data, so the number of channels should be as small as possible. In [19], the authors carry out an investigation over the electrode selection in order to determine the most relevant channels that may maximize the performance of ML models. By using ReliefF and NCA (Neighborhood Component Analysis) methods, there were identified the top 10 channels [19]. The experiments were carried out on DEAP dataset. ReliefF was proposed in [55] and it represents an extension of the RELIEF algorithm [56, 57]. The main advantage of Relief-F is its ability to manage noisy and incomplete data. Proposed in [58], NCA is based on Mahalanobis distance in KNN algorithm. The top 10 channels of DEAP, considered relevant for emotion recognition, are represented in Table 5 [19].

Our goal was to use a small number of channels, so we have selected for our study the following channels: *FP1, AF3, F7, T7, FP2*, as they are the intersection of the algorithms' results.

4.2 Features Extraction

For features extraction, we have used FFT (Fast Fourier Transform) as in [18] with the following changed parameters: number of channels = [1, 2, 4, 8, 17] according to FP1, AF3, F7, T7, FP2 EEG channels and sub-bands theta 4–8 Hz; alpha 8–13 Hz; low beta 13–22 Hz; high beta 22–30 Hz, gamma 30–100 Hz.

4.3 The Selection and Training of the ML Model

According to data presented in Table 3, the 1D-CNN (4 Convolutional + 3 Dense) model is the best performing model using the DEAP dataset for AER [18]. In terms of accuracy, the model 1D-CNN has 99.89%, and 99.83% performance for valence and, respectively, arousal compared with the performance of other models presented in Table 3 with accuracy under 99%. The 1D-CNN has

Table 5 Top 10 channels for emotion recognition [19]

ReliefF	FP1	AF3	F3	F7	T7	O1	Oz	FP2	F8	P8
NCA	FP1	AF3	F7	T7	CP5	P7	FP2	AF4	FC6	T8

four convolutional layers and three dense layers and over 4,381,410 total parameters.

To recognize the seven emotions: *boredom, confusion, frustration, curiosity, excitement, concentration,* and *anxiety*—considered recurrent while learning in the academic environment—we have adapted the aforementioned 1D-CNN model for 5 EEG channels. The number of chosen features is 25 (i.e. 5 EEG channels × 5 sub-bands). So, the total number of parameters decreases more than half, 1,628,898. For the four convolutional and three dense layers, the activation function is ReLU and for the output layer is Softmax. For training, we set batch_size = 100, arbitrary epochs = 100 and implemented early stopping techniques to reduce the training time. Using early stopping techniques, we avoid the overfitting and the training of the models stops after a minimum of 22 epochs. The training of the 14 channels models stops after 28 epochs for anxiety, 25 epochs for boredom, 67 epochs for concentration, 28 for confusion, 33 for curiosity, 39 for excitement, and 65 for frustration. The training of the 10channels-ReliefF models stops after 46 epochs for anxiety, 63 epochs for boredom, 61 epochs for concentration, 33 epochs for confusion, 34 epochs for curiosity, 33 epochs for excitement, and 25 epochs for frustration. Moreover, the training of the 10 channels NCA models stops after 34 epochs for anxiety, 51 epochs for boredom, 46 epochs for concentration, 35 epochs for confusion, 22 epochs for curiosity, 24 epochs for excitement, and 46 for frustration. The training of the five channels models stops after 42 epochs for anxiety, 57 epochs for concentration, 34 for confusion, 55 for curiosity, 34 for excitement, 29 for frustration, and 51 for boredom.

Our model is described in Table 6.

The PAD model has been used for the recognition of the seven above-listed emotions. We have used the mean and standard deviation values provided in [36] for pleasure, arousal and dominance. Since the values range for pleasure, arousal, and dominance in [36] is [−1,1] and in DEAP is [1, 9], we have determined the minimum and maximum values of PAD dimensions in DEAP range for boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety by using the formulas:

$$\text{Min} = (\text{Mean} - \text{Standard Deviation}) * 4 + 5$$

$$\text{Max} = (\text{Mean} + \text{Standard Deviation}) * 4 + 5.$$

where mean and standard deviation values are taken from Table 2. The minimum and maximum PAD values are presented in Table 7.

We have used the values from Table 7 to label DEAP records with 1 or 0, depending on the presence or non-presence of one of the following emotions: *boredom, confusion, frustration, curiosity, excitement, concentration, anxiety*. As an example, we use the following code for boredom recognition and a similar approach is used for the other emotions according to values from Table 7.

```
if (Valence >= 1.64 and Valence<=3.16) and (Arousal >= 1.56 and Arousal<=3.48) and (Dominance >= 2.84
and Dominance<=4.52):
   #presence boredom
   label  = 1
 else:
   #not presence boredom
   label = 0
```

For the evaluation of the model, we have calculated the accuracy, precision, recall and F1 score, because the classes are unbalanced as we can observed in Table 8. Also, the classical confusion matrix is provided for each emotion recognition model.

4.4 Prediction

Using the trained models, we were able to make predictions with respect to emotions recognition. We use the models to perform more experiments to obtain the predictions for various instances of data.

4.5 Explanation of the Predictions—Our Analysis Method Based on LIME

A major disadvantage of LIME algorithm is that the algorithm returns different results at each execution. To explain and to better understand the predicted results, we propose an analysis method based on LIME, derived from ensemble techniques based on majority voting principle.

The steps of our method are:

- LIME is run more times (in our experiments we have run it 20 times).
- For each run, the influence of the features on the prediction is obtained (some features support the presence of

Table 6 The model used for the recognition of the chosen seven emotions

Layer (type)	Output shape	Param #	Connected to
input_1 (InputLayer)	[(None, 25, 1)]	0	[]
conv1d (Conv1D) filters = 32, kernel_size = 6	(None, 25, 32)	224	['input_1[0][0]']
batch_normalization	(None, 25, 32)	128	['conv1d[0][0]']
max_pooling1d pool_size = 2	(None, 12, 32)	0	['batch_normalization[0][0]']
conv1d_1 (Conv1D) filters = 32, kernel_size = 6	(None, 12, 32)	6176	['max_pooling1d[0][0]']
batch_normalization_1	(None, 12, 32)	128	['conv1d_1[0][0]']
max_pooling1d_1 pool_size = 2	(None, 6, 32)	0	['batch_normalization_1[0][0]']
conv1d_2 (Conv1D) filters = 32, kernel_size = 6	(None, 6, 32)	6176	['max_pooling1d_1[0][0]']
batch_normalization_2	(None, 6, 32)	128	['conv1d_2[0][0]']
max_pooling1d_2 pool_size = 2	(None, 3, 32)	0	['batch_normalization_2[0][0]']
concatenate	(None, 21, 32)	0	['max_pooling1d[0][0]', 'max_pooling1d_1[0][0]', 'max_pooling1d_2[0][0]']
conv1d_3 (Conv1D) filters = 128, kernel_size = 6	(None, 21, 128)	24,704	['concatenate[0][0]']
batch_normalization_3	(None, 21, 128)	512	['conv1d_3[0][0]']
max_pooling1d_3 pool_size = 2	(None, 10, 128)	0	['batch_normalization_3[0][0]']
flatten (Flatten)	(None, 1280)	0	['max_pooling1d_3[0][0]']
dense (Dense) units = 1024, activation = 'relu'	(None, 1024)	1,311,744	['flatten[0][0]']
dropout (Dropout) rate = 0.2	(None, 1024)	0	['dense[0][0]']
dense_1 (Dense) units = 256, activation = 'relu'	(None, 256)	262,400	['dropout[0][0]']
dropout_1 (Dropout) rate = 0.2	(None, 256)	0	['dense_1[0][0]']
dense_2 (Dense) units = 64, activation = 'relu'	(None, 64)	16,448	['dropout_1[0][0]']
dropout_2 (Dropout) rate = 0.2	(None, 64)	0	['dense_2[0][0]']
dense_3 (Dense) units = 64, activation = 'softmax'	(None, 2)	130	['dropout_2[0][0]']
Trainable params: 1,628,450 Non-trainable params: 448			

a specific emotion, other features support the absence of that emotion).

- For each feature, the number of occurrences of the feature in the subset of features, which support the presence of a specific emotion, is computed, as well as the number of occurrences of the feature in the subset of features, which support the absence of the that emotion.
- For each feature, the absolute difference of the two frequencies from above is calculated.
- Afterwards, the features are sorted in a descending order, based on their absolute differences; the feature with the highest absolute difference is the most contributing to predict the presence or absence of an emotion. If the number of appearances of the feature in the subset of features which supports the presence of the emotion is higher than the number of appearances of the same feature in the subset of features which supports the absence of the emotion, then we consider that the feature supports the presence of the emotion, otherwise vice versa.

Table 7 Minimum and maximum PAD values in DEAP range for the seven emotions

Emotion	Pleasure		Arousal		Dominance	
	Min	Max	Min	Max	Min	Max
Boredom	1.64	3.16	1.56	3.48	2.84	4.52
Confusion	2.08	3.68	4.92	7.24	2.6	4.84
Frustration	1.72	3.16	5.6	8.56	2.4	4.8
Curiosity	4.68	7.08	6.68	8.28	3.6	6.32
Excitement	6.48	8.48	7.2	8.8	5.36	7.68
Concentration	5.68	7.68	5.04	7.2	5.32	7.8
Anxiety	3.24	6.84	6.12	8.6	3.12	5.68

5 Experiments and Results

Overall, we performed 4 sets of experiments: (1) we used Akter et al. model [18]—1D-CNN for the seven emotions recognition and 14 EEG channels; (2) we used 10 EEG channels provided in [19] with ReliefF algorithm and 1D-CNN model, (3) we used 10 EEG channels provided in [19] with NCA algorithm and 1D-CNN model, and (4) we used 5 EEG channels derived from ReliefF and NCA algorithms and 1D-CNN model. The dataset was split in 80% for training and 20% for test.

Related to the explanations of the results, we ran our analysis method based on LIME for 14 samples.

5.1 14 EEG channels + 1D-CNN

The first experiments aimed to validate the high performance of Akter et al. model [18] for boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety recognition. We used 14 channels (Table 4) and we considered the trained models and their performance as benchmark. In Table 9, we present the obtained performances: the test accuracy is over 99% for all the seven emotions (highest is 99.98% for excitement and lowest is 99.87% for confusion, curiosity and concentration). The F1 score varies from 81.89% for frustration to 99.77% for concentration.

5.2 ReliefF-Based 10 Channels + 1D-CNN Model

In the second series of experiments, we used the model from Table 6 and the 10 EEG channels determined by [19] based on ReliefF algorithm. The channels are presented in Table 5, row 1, and Table 10 shows the performance of the models. The test accuracy obtained are between 99.69% (for the frustration) and 99.87% (for the confusion). The highest value for F1 score is 99.42% for concentration. The running time for training varies from 8421.99 s (over 2 h), in the case of frustration recognition, to 16,821.48 s (about 5 h), in the case of concentration recognition.

Table 8 Numbers of records in each class

Emotion	Training dataset		Test dataset	
	Presence of emotion (label = 1)	Absence of emotion (label = 0)	Presence of emotion (label = 1)	Absence of emotion (label = 0)
Boredom	8400	375,600	2800	125,200
Confusion	13,200	370,800	4400	123,600
Frustration	11,700	372,300	3900	124,100
Curiosity	11,700	372,300	3900	124,100
Excitement	5100	378,900	1700	126,300
Concentration	33,900	350,100	11,300	116,700
Anxiety	10,200	373,800	3400	124,600

Table 9 The performance of the 1D-CNN model (14 channels) for the recognition of the seven emotions

Emotion	Performance (%)				
	Test loss	Test accuracy	Precision	Recall	F1_score
Boredom	0.40	99.91	99.79	98.92	99.36
Confusion	0.46	99.87	98.85	98.76	98.80
Frustration	0.33	99.92	99.45	86.42	91.89
Curiosity	0.48	99.87	99.79	99.51	99.65
Excitement	0.94	99.98	99.24	99.61	99.42
Concentration	0.54	99.87	99.70	99.84	99.77
Anxiety	0.31	99.93	99.29	99.75	99.52

5.3 The NCA-Based 10 Channels + 1D-CNN Model

We employed the 10 EEG channels determined in [19] with NCA, as shown in Table 5, row 2. Their performance is presented in Table 11. There can be observed that all the test accuracies for all seven emotions are over 99.60% and the F1 score values are over 96.92%. The smallest running time is 7011.58 s for the curiosity emotion recognition.

Table 10 The performance for the 1D-CNN model (10 EEG channels—ReliefF) for the recognition of the seven emotions

Emotion	Performance (%)					Running time (seconds)
	Test loss	Test accuracy	Precision	Recall	F1_score	
Boredom	0.56	99.81	97.35	98.27	97.81	12,487.59
Confusion	0.37	99.87	99.79	98.31	99.03	10,117.58
Frustration	1.21	99.69	96.70	98.14	97.41	8421.99
Curiosity	0.51	99.85	98.15	99.40	98.77	10,196.85
Excitement	0.76	99.74	98.77	91.43	94.79	9333.75
Concentration	1.18	99.81	99.45	99.40	99.42	16,821.48
Anxiety	2.43	99.74	97.01	98.05	97.53	13,403.73

Table 11 The performance of the 1D-CNN model (10 EEG channels—NCA) for the recognition of the seven emotions

Emotion	Performance (%)					Running time (seconds)
	Test loss	Test accuracy	Precision	Recall	F1_score	
Boredom	0.28	99.91	99.19	98.68	98.94	14,119.25
Confusion	4.80	99.60	99.09	95.80	96.92	10,057.39
Frustration	0.55	99.84	99.49	97.76	98.61	12,463.57
Curiosity	0.71	99.69	97.97	99.19	98.57	7011.58
Excitement	0.29	99.92	99.21	97.64	98.41	7638.51
Concentration	1.20	99.69	98.71	99.39	99.05	12,405.46
Anxiety	0.66	99.64	99.64	99.17	99.40	9943.37

Table 12 The performance of the 1D-CNN model (5 EEG channels) for the recognition of the seven emotions

Emotion/number of records	Performance (%)					Running time (seconds)
	Test loss	Test accuracy	Precision	Recall	F1_score	
Boredom	0.91	**99.64**	94.62	97.41	95.97	8281.34
Confusion	1.16	**99.70**	99.17	96.19	97.63	6004.65
Frustration	0.98	**99.66**	98.98	95.13	96.97	5690.35
Curiosity	0.97	**99.80**	99.74	96.83	98.24	8675.36
Excitement	0.41	**99.91**	98.74	97.75	98.24	5790.36
Concentration	1.14	**99.70**	99.16	98.98	99.07	9342.58
Anxiety	1.91	**99.21**	95.03	88.96	91.76	6655.65

5.4 5 Channels + 1D-CNN Model

In these experiments, we used 5 EEG channels, namely FP1, AF3, F7, T7, FP2 EEG, and the 1D-CNN model presented in Table 6. Table 12 presents the obtained performance for each emotion as well as the running time values in seconds. The running time are under 9342.58 s for all emotions' cases, and there can be noticed that all the test accuracy values are over 99.21%, The F1 score varies from 91.76%, in the case of anxiety to 99.07%, in the case of concentration.

The confusion matrices for the seven emotions' recognition (5 channels + 1D-CNN model) are presented in Fig. 4. There can be observed the high performance for the model, in the case of boredom the model predicted correctly for 124,883 instances the absence of the emotion from 125,200 instances and for 2662 instances the presence of the emotion from 2800 instances; and so on.

Our findings revealed that data from 5 EEG channels are sufficient to obtain high performance trained models for emotion recognition (Tables 12, 13).

5.5 Comparison Between 14 EEG Channels, ReliefF-Based 10 Channels, NCA-Based 10 Channels, 5 Channels Approaches

In Table 13, there are presented comparatively the accuracies for all cases—14 channels, 10 channels-ReliefF, 10 channels-NCA and five channels. We notice that in the case of five channels the accuracy decreases in 18 situations on average by 0.17. We remark that the greater decrease is by 0.72 and there are four cases in which accuracy increases by 0.1, 0.11, 0.01, and 0.17.

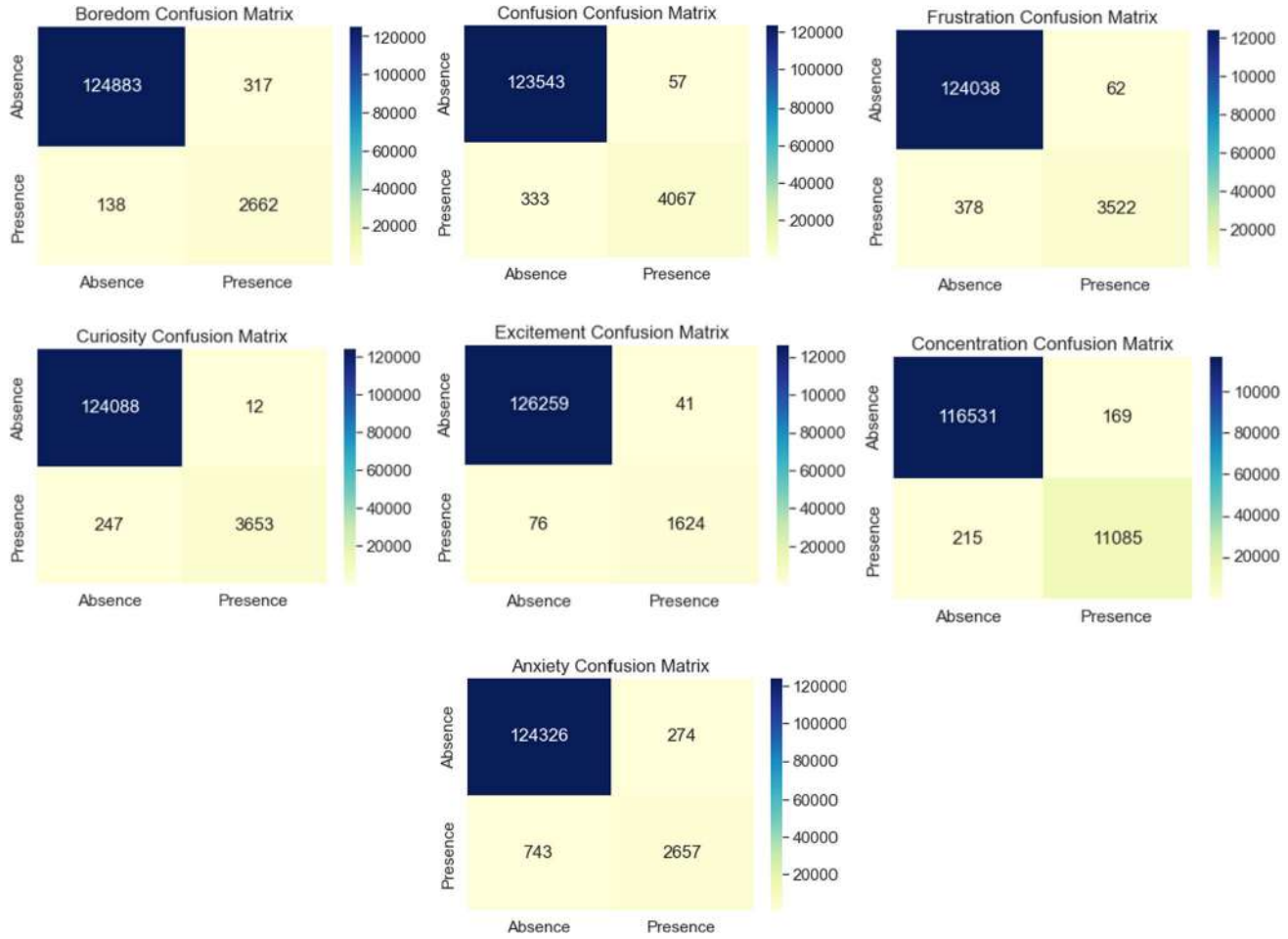

Fig. 4 The confusion matrices for the seven emotions (5 channels + 1D-CNN model)

Table 13 The accuracy for 1D-CNN models

	Boredom	Confusion	Frustration	Curiosity	Excitement	Concentration	Anxiety
14 channels	99.91	99.87	99.92	99.87	99.98	99.87	99.93
10 channels-ReliefF	99.81	99.87	99.69	99.85	99.74	99.81	99.74
10 channels-NCA	99.91	99.60	99.84	99.69	99.92	99.69	99.64
5 channels	**99.64**	**99.70**	**99.66**	**99.80**	**99.91**	**99.70**	**99.21**

In Table 14, there are resumed the F1-score values for all cases—14 channels, 10 channels-ReliefF, 10 channels-NCA and five channels. We notice that in the case of five channels the F1-score decreases in 18 situations on average by 1.84. We remark that the greater decrease is by 7.75 and there are four cases in which values for F-score increases by 0.71, 5.08, 3.45, and 0.02.

The proposed models were implemented on NVIDIA GeForce MX230 GPU with CPU Intel Core i5-10210U @ 1.60 GHz with 8 GB RAM.

Moreover, the running times for training of the models decreases considerably with one exception, as one can notice in Table 15.

In the case of boredom recognition, the training time for 1D-CNN model with five channels decreased by 33% and 41% compared to the training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively. For the confusion, we obtain that the training time for 1D-CNN model with five channels decreased by 40% compared to both training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively. In the case of frustration, the training times for 1D-CNN model with five

Table 14 F1 scores for 1D-CNN models

	Boredom	Confusion	Frustration	Curiosity	Excitement	Concentration	Anxiety
14 channels	99.36	98.80	91.89	99.65	99.42	99.77	99.52
10 channels-ReliefF	97.81	99.03	97.41	98.77	94.79	99.42	97.53
10 channels-NCA	98.94	96.92	98.61	98.57	98.41	99.05	99.40
5 channels	**95.97**	**97.63**	**96.97**	**98.24**	**98.24**	**99.07**	**91.76**

Table 15 Running time for training of the models (seconds)

	Boredom	Confusion	Frustration	Curiosity	Excitement	Concentration	Anxiety
10 channels-ReliefF	12,487.59	10,117.58	8421.99	10,196.85	9333.75	16,821.48	13,403.73
10 channels-NCA	14,119.25	10,057.39	12,463.57	7011.58	7638.51	12,405.46	9943.37
5 channels	**8281.34**	**6004.65**	**5690.35**	**8675.36**	**5790.36**	**9342.58**	**6655.65**

channels decreased by 32% and 54% compared to the training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively. For curiosity, we obtain that the training time for 1D-5 channels model decreased by 14% compared to the training times for the 10 channels—ReliefF model, and it increased by 23% compared to the training time for the 10 channels—NCA model. In the case of excitement, the training times for 1D-CNN model with five channels decreased by 37% and 24% compared to the training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively. For concentration, we obtain the training time for 1D-CNN model with five channels decreased by 44% and 24% compared to the training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively. In the case of anxiety, the training times for 1D-CNN model with five channels decreased by 50% and 33% compared to the training times for the model with 10 channels-ReliefF and 10 channels-NCA, respectively.

5.6 Examples of Applying of LIME-Based Method for Explanations' Generation

Related to the explainability of the model, we ran our analysis method based on LIME for 14 samples.

For each emotion we randomly selected two samples, one for which the model predicted the presence of the emotion (label 1) and one for which the model predicted the absence of the emotion (label 0). We used LIME 20 times for a prediction and for each feature we counted the occurrences in each feature area (the area of features that contribute to the prediction of class 1 and the area of features that contribute to the prediction of class 0). The absolute difference of the two numbers reveals the way in which the value of the feature determines the predicted class. Based on the interpretations presented above, we have several situations:

1. In 20 runs of the LIME algorithm all features contribute to the presence of the emotion (i.e. boredom, confusion, frustration, excitement).
2. In 20 runs of the LIME algorithm some features determine more the presence of an emotion, some determine more the absence of an emotion, while some determine equally the presence and absence of an emotion. By adding the absolute differences for the features from both categories (the presence or absence of an emotion) we obtain the results below.

Below there are presented the results for two samples, one for boredom absence and one for boredom presence. In the first row are the IDs of the features and in the second row are the associated absolute differences. With the orange colour are marked the features which contribute several times to the presence of an emotion and with blue the features which contribute several times to the absence of an emotion. The features with no colours have the absolute difference 0.

Boredom

y_predict = 0 / boredom absence

19	6	23	7	4	11	12	2	5	13	20	24	1	9	14	18	21	0	8	15	16	17	22	3	10
14	12	12	10	8	8	8	6	6	6	6	6	4	4	4	4	4	2	2	2	2	2	2	0	0

In 20 runs:

feature 0 appears 11 times influencing the absence of boredom and 9 times influencing its presence
feature 1 appears 12 times influencing the absence of boredom and 8 times influencing its presence
feature 2 appears 13 times influencing the absence of boredom and 7 times influencing its presence
feature 3 appears 10 times influencing the absence of boredom and 10 times influencing its presence
feature 4 appears 6 times influencing the absence of boredom and 14 times influencing its presence
feature 5 appears 13 times influencing the absence of boredom and 7 times influencing its presence
feature 6 appears 16 times influencing the absence of boredom and 4 times influencing its presence
feature 7 appears 5 times influencing the absence of boredom and 15 times influencing its presence
feature 8 appears 9 times influencing the absence of boredom and 11 times influencing its presence
feature 9 appears 12 times influencing the absence of boredom and 8 times influencing its presence
feature 10 appears 10 times influencing the absence of boredom and 10 times influencing its presence
feature 11 appears 14 times influencing the absence of boredom and 6 times influencing its presence
feature 12 appears 14 times influencing the absence of boredom and 6 times influencing its presence
feature 13 appears 13 times influencing the absence of boredom and 7 times influencing its presence
feature 14 appears 8 times influencing the absence of boredom and 12 times influencing its presence
feature 15 appears 11 times influencing the absence of boredom and 9 times influencing its presence
feature 16 appears 11 times influencing the absence of boredom and 9 times influencing its presence
feature 17 appears 9 times influencing the absence of boredom and 11 times influencing its presence
feature 18 appears 8 times influencing the absence of boredom and 12 times influencing its presence
feature 19 appears 3 times influencing the absence of boredom and 17 times influencing its presence
feature 20 appears 13 times influencing the absence of boredom and 7 times influencing its presence
feature 21 appears 12 times influencing the absence of boredom and 8 times influencing its presence
feature 22 appears 11 times influencing the absence of boredom and 9 times influencing its presence
feature 23 appears 4 times influencing the absence of boredom and 16 times influencing its presence
feature 24 appears 7 times influencing the absence of boredom and 13 times influencing its presence

Summing the absolute differences, we can conclude that the features influence more the absence of boredom (72) than it's presence (62).

y_predict =1 / boredom presence

3	4	23	5	7	13	15	17	19	0	14	18	20	22	9	16	1	6	8	10	11	21	12	2	24
18	18	18	16	16	16	16	16	16	13	12	12	12	12	10	10	9	8	8	8	8	8	6	4	4

In 20 runs:

feature 0 appears 4 times influencing the absence of boredom and 17 times influencing its presence
feature 1 appears 5 times influencing the absence of boredom and 14 times influencing its presence
feature 2 appears 8 times influencing the absence of boredom and 12 times influencing its presence
feature 3 appears 1 times influencing the absence of boredom and 19 times influencing its presence
feature 4 appears 1 times influencing the absence of boredom and 19 times influencing its presence
feature 5 appears 2 times influencing the absence of boredom and 18 times influencing its presence
feature 6 appears 6 time influencing the absence of boredom and 14 time influencing its presence
feature 7 appears 2 times influencing the absence of boredom and 18 times influencing its presence
feature 8 appears 6 times influencing the absence of boredom and 14 times influencing its presence
feature 9 appears 5 times influencing the absence of boredom and 15 times influencing its presence
feature 10 appears 6 times influencing the absence of boredom and 14 times influencing its presence
feature 11 appears 6 times influencing the absence of boredom and 14 times influencing its presence
feature 12 appears 7 times influencing the absence of boredom and 13 times influencing its presence
feature 13 appears 2 times influencing the absence of boredom and 18 times influencing its presence
feature 14 appears 4 times influencing the absence of boredom and 16 times influencing its presence
feature 15 appears 2 times influencing the absence of boredom and 18 times influencing its presence
feature 16 appears 5 times influencing the absence of boredom and 15 times influencing its presence
feature 17 appears 2 times influencing the absence of boredom and 18 times influencing its presence
feature 18 appears 4 times influencing the absence of boredom and 16 times influencing its presence
feature 19 appears 2 time influencing the absence of boredom and 18 time influencing its presence
feature 20 appears 4 times influencing the absence of boredom and 16 times influencing its presence
feature 21 appears 6 times influencing the absence of boredom and 14 times influencing its presence
feature 22 appears 4 times influencing the absence of boredom and 16 times influencing its presence
feature 23 appears 1 time influencing the absence of boredom and 19 times influencing its presence
feature 24 appears 8 times influencing the absence of boredom and 12 times influencing its presence

Summing the absolute differences, we can conclude that the features influence more the presence of boredom (294) than it's absence (0).

The resume of explanations for other 12 samples, for which the absence or the presence of confusion, respectively, frustration, curiosity, excitement, concentration and anxiety are predicted, are shown below.

Confusion

y_predict = 0/confusion absence

12	3	10	16	17	19	24	5	6	8	13	18	2	4	14	15	22	0	1	7	9	11	20	21	23
10	8	8	8	8	8	8	6	6	6	6	6	4	4	4	4	4	2	2	2	2	2	2	2	2

Summing the absolute differences, we can conclude that the features influence more the absence of confusion (94) than it's presence (30).

y_predict = 1/confusion presence

1	8	9	12	13	15	18	19	20	22	23	2	3	4	6	10	11	14	16	21	24	5	17	0	7
20	20	20	20	20	20	20	20	20	20	20	18	18	18	18	18	18	18	18	18	18	16	16	14	14

Summing the absolute differences, we can conclude that the features influence more the presence of confusion (460) than it's presence (0).

Frustration

y_predict = 0/frustration absence

8	6	21	0	18	7	12	1	2	5	9	11	13	17	4	14	16	20	23	19	22	3	10	15	24
16	14	12	10	10	6	6	4	4	4	4	4	4	4	2	2	2	2	2	1	1	0	0	0	0

Summing the absolute differences, we can conclude that the features influence more the absence of frustration (59) than it's presence (56).

y_predict = 1/frustration presence

4	22	8	9	12	15	18	23	2	10	20	6	7	16	19	0	14	17	24	1	3	11	21	5	13
20	20	18	18	18	18	18	18	17	16	16	14	14	14	14	13	12	12	12	10	10	10	10	8	8

Summing the absolute differences, we can conclude that the features influence more the presence of frustration (358) than it's presence (0).

Curiosity

y_predict = 0/curiosity absence

11	21	3	6	1	9	12	18	2	17	23	7	10	13	0	5	14	4	8	15	16	19	20	22	24
17	16	13	11	9	9	9	9	7	7	7	5	5	5	3	3	3	1	1	1	1	1	1	1	1

Summing the absolute differences, we can conclude that the features influence more the absence of curiosity (115) than it's presence (31).

y_predict =1/curiosity presence

6	13	4	19	24	8	10	15	16	20	3	0	5	22	9	17	21	23	7	14	2	11	12	1	18
16	16	14	14	14	12	12	12	12	12	11	10	10	10	9	8	8	8	6	6	2	2	2	0	0

Summing the absolute differences, we can conclude that the features influence more the presence of curiosity (224) than it's presence (2).

Excitement

y_predict = 0/excitement absence

2	7	21	22	14	1	4	5	6	11	17	20	23	3	8	19	9	12	13	15	16	0	10	18	24
14	14	12	12	10	8	6	6	6	6	6	6	6	4	4	4	2	2	2	2	2	0	0	0	0

Summing the absolute differences, we can conclude that the features influence more the absence of excitement (80) than it's presence (54).

y_predict =1/excitement presence

4	10	16	2	6	8	17	18	20	14	22	24	13	19	21	23	9	1	3	15	0	11	5	12	7
20	20	20	18	18	18	18	18	18	17	16	16	14	14	14	14	13	12	12	12	10	10	8	8	4

Summing the absolute differences, we can conclude that the features influence more the presence of excitement (362) than it's presence (0).

Concentration

y_predict = 0/concentration absence

11	16	17	1	8	6	19	20	5	18	9	0	4	7	10	12	14	2	3	13	23	24	15	21	22
20	20	18	16	16	14	10	10	8	8	6	4	4	4	4	4	4	2	2	2	2	2	0	0	0

Summing the absolute differences, we can conclude that the features influence more the absence of concentration (102) than it's presence (78).

y_predict =1/concentration presence

11	6	8	15	16	20	7	0	17	2	4	5	9	12	22	23	3	10	19	21	14	24	1	13	18
20	18	18	16	16	16	14	12	12	10	10	10	10	10	10	10	8	8	8	8	6	6	4	4	0

Summing the absolute differences, we can conclude that the features influence more the presence of concentration (216) than it's presence (48).

Anxiety

y_predict = 0/anxiety absence

8	11	12	16	18	2	1	3	4	22	6	14	17	20	24	7	9	13	15	0	19	21	23	5	10
17	17	17	17	15	13	11	11	9	9	7	7	7	7	6	5	5	5	5	3	3	3	3	1	1

Summing the absolute differences, we can conclude that the features influence more the absence of anxiety (149) than it's presence (55).

y_predict =1/anxiety presence

17	18	20	0	6	10	16	11	21	22	2	5	14	23	24	7	13	19	1	3	12	8	15	4	9
18	16	16	14	14	14	14	12	12	12	10	10	10	10	10	8	8	6	4	4	4	2	2	0	0

Summing the absolute differences, we can conclude that the features influence more the presence of anxiety (210) than it's presence (20).

6 Conclusions

Our interest in realising the presented study lies in building a high performing emotion recognition model based on psychological data, with regard to the emotions often encountered within educational processes and to explain the results obtained by model. Therefore, we selected seven emotions considered relevant in learning: boredom, confusion, frustration, curiosity, excitement, concentration, and anxiety. The emotions were represented using PAD model and the measurements provided by Russel and Mehrabian [36]. We adapted the 1D-CNN model designed by Akter et al. in [18] for 5 EEG channels and obtained a high performing model using only 25 features (i.e. 5 EEG channels $\times$ 5 sub-bands) with the accuracies: boredom—99.64%, confusion—99.70%, frustration—99.66%, curiosity—99.80%, excitement—99.91%, concentration—99.70%, anxiety—99.21%. To explain the predictions, we examined the results with a method based on LIME. Our findings show that running LIME more times and analysing the frequencies of appearances of features supporting the absence or presence of emotion is a solution to explain the outcomes of the model.

The limits of our study are caused by the fact that data used in AER has not been acquired within a learning scenario and that the classes were unbalanced. Designing and conducting experiments to collect EEG data from the participants in the educational process are difficult tasks. However, as we demonstrated that the usage data from 5 EEG channels is enough, the process can be simplified, and we intend to setup such kind of experiments for acquiring data in an educational setting. Another limit of our analysis of the predictions is not taking under consideration the weights of the features generated by LIME. Aware of the limits of our research, we consider that our goal has been achieved, we have obtained a performant AER based on signals from 5 EEG channels and provided a method to explain the predictions.

In our future research, we will take this into account to build a stronger explanation method for AER which includes the weights of the features. At a given moment, a person does not feel a single emotion, but a mixture of emotions. In the presented study, we used binary classification precisely with the idea of capturing the existence of several emotions at the same time. A future direction is to build an emotion recognition model based on multilabel classification through which to specify the spectrum of emotions felt by a person at a given moment.

Author Contributions Conceptualization: G.M., E.G.D., and D.Ș., methodology: G.M., E.G.D., and L.A.I.; software: E.G.D.; writing—original draft: G.M. and D.Ș.; writing—review and editing: G.M., E.G.D., D.Ș.; supervision: G.M.; validation: L.A.I. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by a grant of the Petroleum-Gas University of Ploiesti, project number 11061/2023, within Internal Grant for Scientific Research.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Hwang, G.J., Xie, H., Wah, B.W., Gašević, D.: Vision, challenges, roles and research issues of artificial intelligence in education. Comput. Educ. Artif. Intell. (2020). https://doi.org/10.1016/j.caeai.2020.100001
2. Ouyang, F., Jiao, P.: Artificial intelligence in education: the three paradigms. Comput. Educ. Artif. Intell. **2**, 100020 (2021). https://doi.org/10.1016/j.caeai.2021.100020
3. Pekrun, R.: Emotions and Learning. Educational Practices Series, vol. 24. https://www.iaoed.org/downloads/edu-practices_24_eng.pdf (2014). Accessed 2 January 2024
4. Pekrun, R., Goetz, T., Titz, W., Perry, R.P.: Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research. Educ. Psychol. **37**(2), 91–105 (2002). https://doi.org/10.1207/S15326985EP3702_4
5. Pekrun, R., Lichtenfeld, S., Marsh, H.W., Murayama, K., Goetz, T.: Achievement emotions and academic performance: longitudinal models of reciprocal effects. Child Dev. **88**, 5 (2017). https://doi.org/10.1111/cdev.12704
6. Frenzel, A.C., Daniels, L., Burić, I.: Teacher emotions in the classroom and their implications for students. Educ. Psychol. **56**(4), 250–264 (2021). https://doi.org/10.1080/00461520.2021.1985501
7. Sutton, R.E., Wheatley, K.F.: Teachers' emotions and teaching: a review of the literature and directions for future research. Educ. Psychol. Rev. **15**, 327–358 (2003). https://doi.org/10.1023/A:1026131715856
8. Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., Strohecker, C.: Affective learning—a manifesto. BT Technol. J. **22**, 253 (2004). https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

9. Moise, G., Nicoara, S.E.: Ethical aspects regarding automatic emotion recognition used in online learning environments. In: Caballé, S., Casas-Roma, J., Conesa, J. (eds.) Ethics in Online AI-Based Systems Risks. and Opportunities in Current Technological Trends. Academic Press, Cambridge (2024)
10. European Parliament: Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (2024). Accessed 4 August 2024
11. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). https://doi.org/10.48550/arXiv.1702.08608
12. Bălan, O., Moise, G., Petrescu, L., Moldoveanu, A., Leordeanu, M., Moldoveanu, F.: Emotion classification based on biophysical signals and machine learning techniques. Symmetry **12**, 21 (2020). https://doi.org/10.3390/sym12010021
13. Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P.D., Murugappan, M., Chakole, Y., Acharya, U.R.: Automated emotion recognition: current trends and future perspectives. Comput. Methods Progr. Biomed. **215**, 106646 (2022). https://doi.org/10.1016/j.cmpb.2022.106646
14. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: review of sensors and methods. Sensors **20**, 592 (2020). https://doi.org/10.3390/s20030592
15. Liu, L., Zheng, S., Xu, G., Lin, M.: Cross-domain sentiment aware word embeddings for review sentiment analysis. Int. J. Mach. Learn. Cybern. (2021). https://doi.org/10.1007/s13042-020-01175-7
16. Egger, M., Ley, M., Hanke, S.: Emotion recognition from physiological signal analysis: a review. Electron. Notes Theor. Comput. Sci. (2019). https://doi.org/10.1016/j.entcs.2019.04.009
17. Koelstra, S., Muehl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. **3**, 18–31 (2012). https://doi.org/10.1109/T-AFFC.2011.15
18. Akter, S., Prodhan, R.A., Pias, T.S., Eisenberg, D., Fresneda Fernandez, J.: M1M2: deep-learning-based real-time emotion recognition from neural activity. Sensors **22**, 8467 (2022). https://doi.org/10.3390/s22218467
19. Topic, A., Russo, M., Stella, M., Saric, M.: Emotion recognition using a reduced set of EEG channels based on holographic feature maps. Sensors **22**, 3248 (2022). https://doi.org/10.3390/s22093248
20. Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: Proceedings IEEE International Conference on Advanced Learning Technologies, Madison, WI, USA, pp. 43–46 (2001). https://doi.org/10.1109/ICALT.2001.943850
21. Shen, L., Wang, M., Shen, R.: Affective e-learning: using "emotional" data to improve learning in pervasive learning environment. Educ. Technol. Soc. **12**(2), 176–189 (2009)
22. D'Mello, S., Graesser, A.: AutoTutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. **2**(4), 1–39 (2012). https://doi.org/10.1145/2395123.2395128
23. Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C.: Automatic recognition of student engagement using deep learning and facial expression. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) Machine learning and knowledge discovery in databases. ECML PKDD 2019. Lecture notes in computer science, vol. 11908. Springer, Cham (2020)
24. Chalfoun, P., Chaffar, S.: Predicting the emotional reaction of the learner with a machine learning technique. In: Martinez-Miron, E., Rebolledo-Mendez, G. (eds.) Workshop on Motivational and Affective Issues in ITS. 8th International Conference on ITS 2006, pp. 13–20 (2006)
25. Alyuz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S.E., Arnrich, B., Esme, A.A.: Semi-supervised model personalization for improved detection of learner's emotional engagement. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16), pp. 100–107. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2993148.2993166
26. Kadar, M., Gutiérrez y Restrepo, E., Ferreira, F., Calado, J., Artifice, A., Sarraipa, J., Jardim-Goncalves, R.: Affective computing to enhance emotional sustainability of students in dropout prevention. In: Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI '16), pp. 85–91. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/3019943.3019956
27. D'Errico, F., Paciello, M., De Carolis, B., Vattanid, A., Palestra, G., Anzivino, G.: Cognitive emotions in e-learning processes and their potential relationship with students' academic adjustment. Int. J. Emotion. Educ. **10**(1), 89–111 (2018)
28. Yu, S., Androsov, A., Yan, H., Chen, Y.: Bridging computer and education sciences: a systematic review of automated emotion recognition in online learning environments. Comput. Educ. (2024). https://doi.org/10.1016/j.compedu.2024.105111
29. Fehr, B., Russell, J.A.: Concept of emotion viewed from a prototype perspective. J. Exp. Psychol. Gen. **113**(3), 464–486 (1984). https://doi.org/10.1037/0096-3445.113.3.464
30. Plutchik, R.: The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)
31. Izard, C.E.: Patterns of Emotions. Academic Press, New York (1972)
32. Ekman, P., Sorenson, E.R., Friesen, W.V.: Pan-cultural elements in facial displays of emotions. Science **164**, 86–88 (1969)
33. Izard, C.E.: Human Emotions. Plenum Press, New York (1977)
34. Ekman, P.: Basic emotions. In: Dalgleish, T., Power, M. (eds.) Handbook of Cognition and Emotion. John Wiley & Sons Ltd, Hoboken, NJ, USA (1999)
35. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
36. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. J. Res. Pers. **11**, 273–294 (1977). https://doi.org/10.1016/0092-6566(77)90037-X
37. Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genet. Soc. Gen. Psychol. Monogr. **121**, 339–361 (1995)
38. Mehrabian, A.: Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr. Psychol. **14**, 261–292 (1996). https://doi.org/10.1007/BF02686918
39. Akhand, M.A.H., Maria, M.A., Kamal, M.A.S., Murase, K.: Improved EEG-based emotion recognition through information enhancement in connectivity feature map. Sci. Rep. **13**, 13804 (2023). https://doi.org/10.1038/s41598-023-40786-2
40. Liu, X., Li, T., Tang, C., Xu, T., Chen, P., Bezerianos, A., Wang, H.: Emotion recognition and dynamic functional connectivity analysis based on EEG. IEEE Access **7**, 143293–143302 (2019). https://doi.org/10.1109/ACCESS.2019.2945059
41. Rahman, M.M., Sarkar, A.K., Hossain, M.A., Hossain, M.S., Islam, M.R., Hossain, M.B., Quinn, J.M.W., Moni, M.A.: Recognition of human emotions using EEG signals: a review. Comput. Biol. Med. **136**, 104696 (2021). https://doi.org/10.1016/j.compbiomed.2021.104696

42. Wang, J., Wang, M.: Review of the emotional feature extraction and classification using EEG signals. Cogn. Robot. **1**, 29–40 (2021). https://doi.org/10.1016/j.cogr.2021.04.001
43. Islam, Md.R., Islam, Md.M., Rahman, M.M., Mondal, C., Singha, S.K., Ahmad, M., Awal, A., Islam, M.S., Moni, M.A.: EEG channel correlation based model for emotion recognition. Comput. Biol. Med. **136**, 104757 (2021). https://doi.org/10.1016/j.compbiomed.2021.104757
44. Cao, S., Liu, H., Hou, Z., Li, X., Wu, Z.: EEG-based hardware-oriented lightweight 1D-CNN emotion classifier. In: 15th International Conference on Intelligent Human–Machine Systems and Cybernetics (IHMSC), Hangzhou, China, pp. 210–213 (2023). https://doi.org/10.1109/IHMSC58761.2023.00056
45. Akhand, M.A.H., Maria, M.A., Kamal, M.A.S., Shimamura, T.: Emotion recognition from EEG signal enhancing feature map using partial mutual information. Biomed. Signal Process. Control **88**(Part A), 105691 (2024). https://doi.org/10.1016/j.bspc.2023.105691
46. Li, F., Hao, K., Wei, B., Hao, L., Ren, L.: MS-FTSCNN: an EEG emotion recognition method from the combination of multi-domain features. Biomed. Signal Process. Control (2024). https://doi.org/10.1016/j.bspc.2023.105690
47. Huang, W., Chen, Y., Jiang, X., Zhang, Z., Chen, Q.: GJFusion: a channel level correlation construction method for multimodal physiological signal fusion. ACM Trans. Multimed. Comput. Commun. Appl. (2023). https://doi.org/10.1145/3617503
48. Zheng, W., Pan, B.: A spatiotemporal symmetrical transformer structure for EEG emotion recognition. Biomed. Signal Process. Control (2024). https://doi.org/10.1016/j.bspc.2023.105487
49. Fan, F., Xie, H., Tao, J., Li, Y., Pei, G., Li, T., Lv, Z.: ICaps-ResLSTM: improved capsule network and residual LSTM for EEG emotion recognition. Biomed. Signal Process. Control (2024). https://doi.org/10.1016/j.bspc.2023.105422
50. Scott, A.C., Clancey, W.J., Davis, R., Shortliffe, E.H.: Explanation capabilities of production-based consultation systems. Am. J. Comput. Linguist. **14**, 1–50 (1977)
51. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
52. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. U.S.A. **116**(4), 22071–22080 (2019). https://doi.org/10.1073/pnas.1900654116
53. Guegan, D.: A Note on the Interpretability of Machine Learning Algorithms. https://shs.hal.science/halshs-02900929 (2020). Accessed 2 May 2024
54. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**, 832 (2019). https://doi.org/10.3390/electronics8080832
55. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) Machine Learning: ECML-94. ECML 1994. Lecture Notes in Computer Science, vol. 784. Springer, Berlin, Heidelberg (1994)
56. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Sleeman, D., Edwards, P. (eds.) Machine learning proceedings, pp. 249–256. Morgan Kaufmann, USA (1992)
57. Kira K., Rendell, L.A.: The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the tenth national conference on Artificial Intelligence (AAAI'92), pp 129–134. AAAI Press, USA (1992)
58. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Proceedings of the 17th International Conference on Neural Information Processing Systems (NIPS'04), pp. 513–520. MIT Press, Cambridge, MA, USA (2004)

Towards Building Creative Collaborative Learning Groups Using Reinforcement Learning

Monica Vladoiu *monica@unde.ro*
UPG University of Ploiesti
Ploiesti, Romania

Gabriela Moise *gmoise@upg-ploiesti.ro*
UPG University of Ploiesti
Ploiesti, Romania

Zoran Constantinescu *zoran@unde.ro*
UPG University of Ploiesti
Ploiesti, Romania

Abstract

Increasing creative skills in collaborative groups is of huge interest for stakeholders in education, industry, policy making etc. However, construction of "the most" creative groups given a cohort of people and a set of common goals and tasks to perform is challenging. The complexity of this undertaking is amplified by the necessity to first understand and then measure what "the most" creative means in a particular situation. We present here our method of semi-automatic building of "the most" creative learning groups given a cohort of students and a particular learning context based on *reinforcement learning* (an adapted Q-learning algorithm). Various attributes that influence individual and group creativity may be considered. A case study on using this method with our Computer Science students is also included. However, the method is general and can be used for building collaborative groups in any situation, with the appropriate "the most" creative goal and attributes.

Keywords: Collaborative Groups, Optimally Creative Learning Groups, Reinforcement Learning, Computer Supported Collaborative Learning

1. Introduction

Educational paradigms adjust continuously to stay tuned with the continuous change in our society. Promoting collaboration and boosting creativity in learning are major trends nowadays. Hence, increasing creative and collaborative skills of both students and employees is currently of huge interest for stakeholders in education, industry, policy making, etc. However, creativity is a concept still highly debated in the psychological literature. Sternberg et al. see *creativity as the ability to produce work that is novel (i.e., original, unexpected), high in quality, and appropriate* [19]. *Group or collective creativity* is a much more recent topic in the literature, which takes into account *the social nature of the creative act* [7]. Nevertheless, group creativity means much more that summing up the individual creativities of group members, as the interactions that take place between them, the stimulation, both cognitive and motivational, that results from these interactions, the diversity of their backgrounds, their abilities and knowledge contribute further to adding value in creative processes, resulting in *a true synergy* [5]. *Collaborative learning groups* are working groups that evolve during common educational scenarios that unfold over long periods of time and, generally, become teams, based on the evolution of the relationships inside the group. Their

creativity can be approached within augmented collaborative learning environments, in which group members work creatively, both individually and collaboratively, to fulfill particular tasks, to complete specific projects, or to achieve particular goals. The results of their work can be problem solutions, papers, overviews, (pieces of) software or hardware, documents, essays etc. These results are evaluated by instructors who assess the creativity of the resulted products and, this way, a measurement of group creativity can be obtained. An example of an augmented collaborative learning environment can be a classroom with instructional materials and/or equipments (e.g. drawings, robots, drones, maps etc.), along with a set of teaching and learning methods (problem-based learning, brainstorming, project-based learning, game-based learning, etc.) that stimulate innovation and imagination.

Various approaches may be taken to build optimally creative collaborative learning groups given a cohort of students and a learning context. During the eighties, Amabile has developed *The Componential Model of Creativity* for individual creativity, which she has further extended to team creativity and innovation in organizations [1], [4]. Further, in 2012, she proposed a componential theory of creativity, which includes three within-individual components (domain-relevant skills, creativity-relevant processes, task motivation) and a non-individual component, i.e. the social environment [2, 3]. Her theory points out that creativity calls for a convergence of all these components and that creativity should be at peak when a deeply motivated and very skillful in creative thinking person with high domain expertise works in an environment providing highly for creativity [2, 3]. Similarly, Taggar has shown that team creativity is significantly influenced by relevant processes that emerge as part of group interaction [18]. Further, based on the theoretical bases of synergy, in [5], the authors identify the cognitive, social, and motivational factors that influence the increase of group creativity: exchange of ideas, potential for competitiveness that allow individuals to compare their performances with the ones of their teammates, concept, product and perspective sharing, intrinsic motivation, openness to new experiences, etc.

Contextual factors that influence group creativity are summarized in [21] as being factors that facilitate team creativity (supervisory and co-workers support, psychological safety, group process), factors that obstruct generation of creative ideas (conformity, insufficient resources, bureaucratic structure), and uncertain factors (team diversity, conflicts in teams, group cohesion). In [6], the authors analyzed the cause-effect relationships between 6 factors: team creativity, exploitation, exploration, organizational learning culture, knowledge sharing, and expertise heterogeneity. Several correlations have been found, for example, to sustain high levels of team creativity both organizational learning culture and knowledge sharing should be high. A model of collaborative creativity that takes into account four categories of variables and three categories of processes which influence creativity and innovation is provided in [13]. The four categories of variables are: group member variables, group structure, group climate, and external demands, while the three categories of processes are: cognitive, motivational, and social. The research in [15] shows that creativity is multifaceted and it can be assessed by measuring *fluency* (creative production of non-redundant ideas, insights, problem solutions, or products), *originality* (uncommonness or rarity of these outcomes), and *flexibility* (how creativity manifests itself when using comprehensive cognitive categories and perspectives).

However, construction of creative groups is not straightforward and, up to our knowledge, research on this subject is rather scarce. An overview is available in our previous works [11, 12], though most of the (very loosely) related work do not use data mining techniques, machine learning, nor intelligent data analysis neither take into account individual creativity measures to support the construction of creative collaborative groups.

We introduce here *a method based on reinforcement learning* (an adapted Q-learning algorithm) *to semi-automatically build optimally ("the most") creative learning groups*, given a cohort of students and a particular learning context. Various attributes that influence individual and group creativity may be considered. However, the method is general and can be used for obtaining "the most" creative groups in any learning, working, or other collaborative situation. Reinforcement learning is an area of machine learning concerned with how software agents learn to take actions within an environment (as a result of their

interaction with that environment) so that they maximize some cumulative reward. In the typical reinforcement learning model, an agent is connected to its environment via perception and action. On each step of its interaction with the environment, a particular agent receives as input some indication of the current state of this environment and it then chooses an action that changes the state of the environment. The value of this state transition is communicated to the agent through a scalar reinforcement signal. The agent is expected to behave by choosing actions that tend to increase the long-run sum of values of this reinforcement signal. It can learn to do this in time by prearranged trial and error iterations directed by a wide variety of algorithms [9]. The most well-known algorithms for solving problems using reinforcement learning are based on Q-learning [20] and SARSA-learning [16].

During this work, we used an adapted Q-learning algorithm to build "the most" (optimally) creative groups given a cohort of students and a particular learning context. *Individual creativity* and *motivation* are the attributes that influence group creativity, which have been taken into account in the case study included here. Individual creativity has been assessed using the Gough Creative Personality Scale [8], while students' motivation has been determined using our adapted questionnaire based on MSLQ [14]. This case study has been performed with our Computer Science students and it is based on the algorithm introduced briefly in [12]. Particularly, we have determined, for each student, to what group's creativity s/he would contribute the most, given the attributes considered.

The structure of the paper is as follows: the general Q-learning algorithm is shown briefly in Section 2, while the third one includes the adapted version used in our method for building creative groups. Section 4 presents the results obtained when using this method in a particular educational context, while the last section includes some conclusions and future work ideas.

2. The Q-Learning Algorithm

In brief, the Q-learning algorithm is a reward learning algorithm that starts with an initial estimate *Q(s, a)* for each pair *<state, action>*. When a certain action "a" is chosen in a state "s", the intelligent system gets a reward R(s, a) and the next state of the system is acknowledged. The Q-learning algorithm estimates the function value-state-action as follows:

$$Q(s,a) \coloneqq Q(s,a) + \alpha\big(R(s,a) + \gamma max_{a'} Q(s',a') - Q(s,a)\big) \quad (1)$$

Where $\alpha \in (0,1)$ is the learning rate, $\gamma \in (0,1)$ is the discount factor, and s' is the state reached after executing the action "a" in the state "s". Values for the learning rate and for the discount factor are selected according to [10]. The higher the value of the learning rate the faster the learning is, while a value of 0 means that the value for Q is never updated, and therefore the system never learns. When the learning rate is 1 it means that the immediate reward is much more important than a past reward. A balance between the immediate rewards and the past rewards is sought for in dynamic environments. In our first experiments, we had used a 0.5 learning rate. The discount factor takes values between 0 and 1. Closeness to 1 means that a future reward is more important than an immediate reward, i.e. that the importance of a future reward is significant (as γ is still below 1). The pseudo-code of the Q-learning algorithm is presented below [20].

```
initialize random Q-values (Q(s, a)) for each pair <state, action>
      repeat (for each scenario)
            initialize s
            repeat (for each step of scenario)
                  choose a using a policy derived from Q
                  observe s, execute a, observe reward R, observe s'
                  update Q(s,a)
                  s:=s'
            until s is terminal
```

3. GC-Q-Learning Algorithm for Building Creative Groups

The GC-Q-Learning adapted algorithm used in our method is presented further on. It starts with *n* students. For each student, *a creativity vector c* that includes "m" individual characteristics that influence creativity is constructed, i.e. $c=(c_1, c_2, ..., c_m)$. In fact, for this algorithm, a student is not a particular person, but a particular *type of student* given by her set of characteristics. Therefore, all the students having the same creativity vector will be a generic student for our algorithm. *A state* consists of this creativity vector and the group number of each student, while *an action* consists in moving a student to another group in which *he would contribute the most to increasing group creativity*. *Q* expresses the quality of association between a state and an action. Our goal is to build "the most" creative *k* groups (*k* being given). The state space includes the set of tuples that can be built taking into account that each characteristic can have a finite number of values. The size of action space is given by the number of groups (*k*) to be constructed. More individual characteristics taken into consideration would lead to increasing the dimension of the state space, which can result in difficulties in implementing the algorithm. For the time being, the main characteristics taken into account in our case studies have been the following: *individual creativity, motivation, domain knowledge,* and *inter-personal affinities* [11, 12].

When using this algorithm, a large number of students to be grouped at once can be challenging as well. Thus, for *n* students and *k* groups, each group will contain the closest natural number larger or equal with *n/k* (when *n/k* is not a natural number, the rest of students (*m*) is distributed, randomly, one student per each of the already formed groups). The number of groups that can be obtained this way is $C_n^{n/k}$, which can be very large for particular values of *n* and *k*. Though, for reasonable group size, between 10-30 students, the algorithm can be applied easily, while significantly larger cohorts of students need to be divided in smaller groups, randomly, and only then perform the algorithm on these groups.

The GC-Q-Learning algorithm computes the best organization of a cohort of students in creative groups, while the environment consists mainly of this *structural organization* [17]. Of course, the structure of the groups generally changes over time, as the system learns from its interactions with its environment how to construct more and more creative groups. The reward is the value of group creativity and it ranges between 1 and 5. The global creativity objective is to obtain a final state, *namely an organization of students in groups*, in which either each group will have a creativity value larger than a desired threshold or the average creativity on all the groups will be higher than such a threshold. The GC-Q-learning adapted algorithm is as follows:

1. Build a bi-dimensional matrix *Q* for all the possible pairs <*state, action*>. The columns consist of $(c_1, c_2, ..., c_m, no\_group, action\_number, q)$. A value of the *action_number* of i means that if a particular type of student (given by his creativity vector $c_1, c_2, ..., c_m$) will be moved to the group having the value of *no_group* i then her contribution to group creativity is quantified by q (in this stage). All the elements in the *q* column may be initialized with 0 or with a randomly chosen low value. On each line of the matrix, the data that corresponds to each type of student involved in the grouping process is included, i.e. the values of his characteristics, the current group number, the action number, and the value computed for *q (that quantifies a potential for creativity)*. One particular type of student could have more corresponding lines, one for each combination <*current group number, action*>;
2. Initialize the *optimal_policy* with an initial policy. In our case, the optimal policy is an optimal grouping of students that maximizes group creativity. The initial grouping is set by the instructor and the students together and experience shows that they tend to group as cliques based on their inter-personal affinities;
3. Group the students and have them carry on working sessions (in the case study presented here those were several *online brainstorming* sessions, but any collaborative situation involving creativity can be used), in which each group's

creativity is assessed and its score is assigned to the reward *R(s,a)*. The values of *R(s,a)* are obtained with help from human experts (in our tests, they have scored each idea in each session). We may say that R *materializes* that potential for creativity (q). Then, the matrix *Q* is re-calculated for each such working session. This procedure is presented below.

```
procedure working_session_computation
select action of (optimal_policy)  /* student grouping*/
compute R(s,a)
compute table Q /* using formula (1)*/
```

4. Analyze the group creativity for each group against the global objective (the optimal grouping policy), which is getting closer to the maximum value possible for R, for each group or for all the groups. Re-iterate from step 3, if necessary.

Once the optimal policy consisting in tuples *(c_1, c_2, …, c_m, group number)* is obtained, an intelligent system (or an agent) based on this algorithm has learned to build the most creative groups given the circumstances. Consequently, it can make prediction for *each new type of student*, given his set of characteristics, using advanced classification techniques (Bayesian networks, neural networks etc.). The predictions consist of a series of group numbers, which are presented sorted decreasingly according to the contribution made by that particular generic student to each group's creativity. Thus, the first number in the series is of the group in which that generic student would contribute the most to the group creativity, the second one of the group in which she would make the second best contribution, and so on. We have already worked on this idea of building the most creative and innovative collaborative groups using Bayes classifiers with encouraging results [11].

4. Experimenting with the GC-Q-Learning Algorithm

In this section, the data obtained during one of our testing of the GC-Q-Learning algorithm is presented. This particular one was performed on 36 undergraduates in Computer Science, who participated voluntarily in three working sessions. We have grouped and re-grouped these students during the three sessions, aiming at having each student being a member of the group to which creativity s/he contributes "the most" according to our assumptions. In this testing, *individual creativity* and *motivation* were the attributes included in the creativity vector. The Gough Creative Personality Scale [8] has been used to assess each individual's creativity. Generally, the Gough Score values range between -12 and 18. Students' motivation has been determined using our adapted questionnaire based on MSLQ (Motivated Strategies for Learning Questionnaire) [14] (both are presented in the apendices). It contains 31 statements with a possible value between 1 and 7 (1 means that the statement is totally untrue, 7 means that the statement is completely true, while scores between 2 and 6 are somewhere in between). In our trials, we considered low motivation between 31 and 93 (the associated motivation score being 0), medium motivation between 94 and 155 (motivation score 1), and high motivation between 156 and 217 (motivation score 2). After evaluation, we have obtained the following classification of students with respect with their creativity vector:

Table 1. Classification of students with respect with the creativity vector.

Creativity vector ***(Individual Creativity, Motivation)***	**Number of students**
(2,1)	6
(2,2)	3
(3,1)	9
(3,2)	12
(4,1)	6

The students regrouped repeatedly in groups of four by permutation. Three online brainstorming sessions took place on subjects of interest for them: (1) the improvement of both the curricula and the syllabuses for our Computer Science programs (undergraduate and graduate), (2) the preferred teaching and learning methods, and (3) the enhancement of their student life within university and campus alike. Each session had to end with a final conclusion on the issues discussed. We used brainstorming here just for measuring group creativity, but any other way of appropriate evaluation can be used. These sessions have taken place online to avoid some of the shortcomings of the face-to-face brainstorming sessions emphasized in the literature.

In total, the creativity for the 27 groups (three sessions, each session involved nine groups) has been measured using the scores below (human expert evaluated):

- A score R1 has been given after evaluation of the quality of ideas generated by each group of students;
- A score R2 has been given for the frequency of ideas generated by each group of students;
- A score R3 has been obtained for the quality of the final conclusion of each session; this evaluation was performed by human experts.
- A final score, R, has been computed as the mathematical mean of the three scores above. It will be the reward used by the algorithm (Table 2).

For this working session, the Q matrix had 135 lines (because there are 5 types of students having the characteristics (3,1), (3,2), (2,1), (2,2) and (4,1) and 27 groups) and 4 columns. Each column consists in, respectively, the Gough score, the motivation value, the action number (that means to move her in the group in which she would contribute the most to that group's creativity, if included in it, given her characteristics), and the q value. On each line of the matrix the data that correspond to each type of student involved in the grouping process is available, i. e. the values for: the Gough score, the motivation, the current group number, the action number, and the value computed for q. We present below some data sample consisting of 9 groups of 4 (type of) students given by their creativity vector (Table 2: Label C = individual creativity, Label M = motivation score).

Table 2. Creativity Vector (Individual Creativity, Motivation) of each student of each group.

	Student i		**Student j**		**Student k**		**Student l**		
No. of group	**C**	**M**	**C**	**M**	**C**	**M**	**C**	**M**	**R (Reward)**
1	3	1	3	2	3	1	3	1	3
2	2	1	2	2	2	1	3	2	4
3	4	1	4	1	3	2	3	2	4,33
4	3	2	2	1	3	2	3	2	3
5	3	1	2	1	3	2	4	1	3,66
6	3	1	3	1	2	2	4	1	2,66
7	3	1	3	1	2	1	4	1	3,33
8	3	2	2	1	3	2	3	2	3,33
9	3	1	2	2	4	1	3	2	3,33

We present below some testing results obtained while trying to group, in increasingly creative teams, several pools of students having various values for *the creativity vector (Gough score, motivation value)*. In the case study presented further on, we had 5 types of students characteristic-wise, with the above mentioned pairs as follows: (3,1), (3,2), (2,1), (2,2), and (4,1), and we studied 27 possible groups, each formed with 4 students. The value of both the learning rate α and the discount factor γ were 0.5. In Table 3 and Fig. 1, some of the sample data for the students having the creativity vector (3,1) are shown. The action to be performed is moving such a student in a particular group, the computed q value being shown as well. The interpretation of this data snapshot is that a student with the pair (3,1) would

contribute the most to the group creativity if s/he would be a part of group 7, and decreasingly - of group 1, 6, 5, or 9. Group number 7 is composed of 4 students with the characteristics pairs as follows: (3,1), (3,1), (2,1), and (4,1) (according to Table 2).

To use this method, one needs to group the students randomly or based on their interpersonal affinities, then have them work as groups in a learning scenario. Based on the values of their creativity vector and using the adapted Q-learning algorithm, the composition of the groups may change. Thus, a student may be moved to a group for which his q value is among first 30% in decreasing order (*to raise the potential for increasing group creativity*). Then the collaborative creative activity takes place, in our case a second online brainstorming session. Further on, the obtained data (group creativity is the reward of the algorithm) is fed back to the algorithm and, this way, *it learns over time what is the best option of moving a (particular type of) student in the group in which s/he has the maximum contribution to the group's creativity*. The goal here is to obtain a final state, namely *an organization of students in groups*, in which either *each such group has a creativity value larger than a desired threshold or the average creativity on all the groups is higher than such a threshold*.

Table 3. Sample Data for Students with Creativity Vector (3,1).

Gough score	Student motivation	Action: *move to group no*	Q value
3	1	1	3.46875
3	1	2	0
3	1	3	0
3	1	4	0
3	1	5	2.697188
3	1	6	3.295781
3	1	7	**3.798281**
3	1	8	0
3	1	9	2.532188

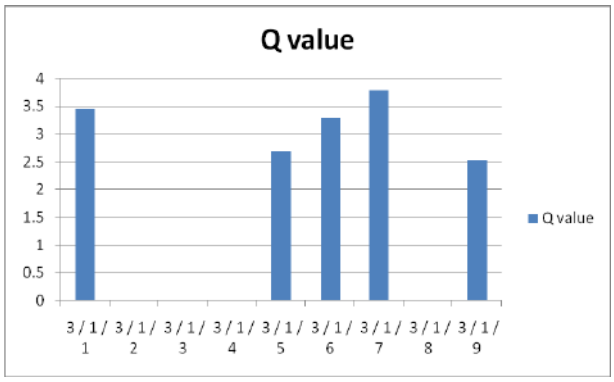

Fig. 1. Q value – students with (Gough, motivation): (3,1).

However, the students are not grouped and re-grouped indefinitely, as the algorithm learns during time in which group a student should be to contribute the most to group's creativity. So, it can make a recommendation in this sense (which, of course, can be followed or not by the instructors and students based on their learning objectives at that time).

We present here some evaluation data obtained during the educational activities related to the Software Engineering class. The performance of the groups is measured by two grades, which are granted based on several criteria that measure both the performance of each group as a whole and each individual contribution. These criteria assess the developed software, the related documentation, the difficulty of the problem, the creative and innovative solutions used during development and for the presentation of the final product, the complexity of the algorithms, the cost-effectiveness of the solution, the degree of being a close-knit team, and so on. The two grades are midterm and final, they being obtained for the initial, respectively, optimally created groups. As it can be seen below in Table 4, the performance of the majority

of students is higher after this collaborative learning experience (the average grade of each group is presented). And this is not just an isolated situation, as we have already performed this kind of grouping, in similar circumstances, for 4 years now, and the results are consistent and show increased learning with respect to both domain expertise and soft skills achieved. Thus, during our work with the students involved, throughout their university years, both as undergraduate and graduate, we have evaluated the creativity of the teams obtained in this way and the results show that they are, indeed, more creative than ad-hoc or buddy teams, as they consistently obtain better evaluations of teamwork results [11, 12].

Table 4. Sample data obtained while evaluating the method.

Year	Team	Midterm average group grade	Final average group grade
2018	1 (8 members)	6.87	7.50
	2 (8 members)	7.00	7.25
	3 (6 members)	9.00	9.33
	4 (7 members)	7.00	7.28
2017	1 (7 members)	6.14	6.85
	2 (5 members)	7.20	7.60
	3 (5 members)	5.00	5.00
	4 (6 members)	8.00	8.33
2016	1 (6 members)	7.16	9.16
	2 (6 members)	6.16	8.16
	3 (4 members)	6.50	7.50
	4 (5 members)	6.40	7.40
2015	1 (6 members)	7.66	9.16
	2 (4 members)	5.00	6.00
	3 (7 members)	8.42	10.00
	4 (3 members)	6.00	7.00

5. Conclusions and Future Work

One of the invariants of nowadays life is, paradoxically, continuous change that takes place in more and more aspects of our life. To keep pace, existent paradigms have to perpetually shift to better adapt to our constantly changing world. In this sense, education and collaboration among people have had an astonishing entwined evolution that allows better accomplishment of important common goals. For example, creativity and innovation are very much valued and sought after both in collaborative learning and collaborative working, as increasing the efficiency and effectiveness of groups of individuals performing together specific activities to achieve common goals, in given contexts, is of crucial importance. Consequently, promoting collaboration and boosting creativity in learning and working are major trends nowadays, so group creativity has become an active topic of creativity research. However, despite the consensus that both individual creativity characteristics and inner interactions inside groups influence collaborative creativity, the construction of “the most” (optimally) creative groups given a cohort of people and a collaborative context is challenging. Various approaches may be taken based on various factors that influence creativity, both at individual and group level. Our approach in this work has been based on two important such factors, namely individual creativity and motivation. Well-known scales have been used as such or adapted to evaluate these factors in case of a cohort of Computer Science undergraduates, who volunteered to participate in this experiment that aimed at increasing group creativity in a collaborative learning context.

During successive online brainstorming sessions we have grouped and re-grouped the participants according with the results provided by the reinforcement algorithm aiming at obtaining “the most” creative groups possible given that particular cohort of students, their evaluated creativity scores, and the particular learning context. The algorithm has learnt, in

time, in which particular group each (type of) student should be, so that s/he can contribute the most to a particular collaborative creativity.

This is work in progress and many future work directions unfold. More experiments on various learning scenarios need to be considered in Computer Science education, as well as in other domains, with diverse cohorts of students, evaluating group creativity using various metrics, maybe using control groups if this can be done respecting the principle of pedagogical fairness, etc. More factors need to be taken into account too, for example, group interactions and the way they develop over time. Testing the method in other collaborative contexts would be valuable as well. Development of a software tool that implements the method would be very useful to assist the construction of the most creative groups in particular collaborative scenarios.

Despite the promising results so far, the method is not to be used exclusively because it has an important limitation, i.e. the fact that all the factors that influence creativity need to be evaluated by numbers, while it is well known that same cannot be assessed that way whatsoever (for example, interpersonal affinities). Combining this method with others that allow using linguistic values, such as weak, strong, etc., seems to provide for a viable solution of semi-automatic grouping people in the most creative groups possible in a given collaborative context, this being the most important future work direction.

Of course, it makes more sense to apply this semi-automatic grouping method for groups of people aiming at becoming teams, during long periods of time, such as university or working years. However, the method can be used also for groups formed for shorter periods of time because it is based on characteristics that quite often have the same values for different people (for example, the creativity vector <individual creativity, motivation>), so the process does not need to start from scratch each time, but just build up on previous results.

References

1. Amabile, T. M.: A Model of Creativity and Innovation in Organizations. Research in organizational behavior. In Staw, B. M., Cummings, L. L. (eds.), Research in organizational behavior, 10, 123-167. Greenwich, CT: JAI Press (1988)
2. Amabile, T. M.: Componential Theory of Creativity. In: Kessler, E. H., (ed.). Encyclopedia of Management Theory. pp. 135-140. SAGE Publications Inc. (2013)
3. Amabile, T. M.: Componential Theory of Creativity. Working paper, http://www.hbs.edu/faculty/Publication%20Files/12-096.pdf, Accessed April 13 2018
4. Amabile, T. M.: Social Psychology of Creativity: A Componential Conceptualization. Journal of Personality and Social Psychology 45(2), 357-377 (1983)
5. Baruah, J., Paulus, P. B.: Enhancing Group Creativity: The Search for Synergy. In E. A. Mannix, J. A. Goncalo & M. A. Neale (eds.), Creativity in Groups (Research on Managing Groups and Teams), 12, 29-56. Emerald Group Publishing Limited (2009)
6. Choi, D. Y., Lee, K. C., Seo, Y. W.: Scenario-Based Management of Team Creativity in Sensitivity Contexts: An Approach with a General Bayesian Network. In: Kun Chang Lee (eds.), Digital Creativity, Individuals, Groups, and Organizations, Integrated Series in Information Systems, 32, 9-113. Springer, New York, (2013)
7. Dictionary of creativity (2018), http://creativity.netslova.ru/Group_ creativity.html. Accessed June 26, 2018
8. Gough. H. G.: A Creative Personality Scale for the Adjective Check List, Journal of Personality and Social Psychology 37, 1398-1405 (1979)
9. Kaelbling, L. P., Littman, M. L., Moore, A. W.: Reinforcement Learning A Survey. Journal of Artificial Intelligence Research 4 (1), 237-285. AI Access Foundation and Morgan Kaufmann Publishers (1996)
10. Leon, F., Şova, I. and Gâlea, D.: Reinforcement Learning Strategies for Intelligent Agents. In Proceeding of the 8th International Symposium on Automatic Control and Computer Science. Iaşi (2004)

11. Moise G., Vladoiu M., Constantinescu Z.: Building the Most Creative and Innovative Collaborative Groups Using Bayes Classifiers. In: Panetto H. et al. (eds) On the Move to Meaningful Internet Systems. OTM 2017 Conferences. Lecture Notes in Computer Science, vol 10573, pp. 271-283. Springer, Cham (2017).
12. Moise, G., Vladoiu, M., Constantinescu, Z.: GC-MAS - a Multiagent System for Building Creative Groups used in Computer Supported Collaborative Learning. In: 8th International KES Conference on Agents and Multi-agent Systems – Technologies and Applications, Advances in Intelligent Systems and Computing, 296, pp. 313-323. Springer, Cham (2014).
13. Paulus, P. B., Dzindolet, M.: Social influence, creativity and innovation. Social Influence 3, 228–247. Taylor & Francis (2008)
14. Pintrich, P. R., Smith, D. A., Garcia, T., Mckeachie, W. J.: Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educational and Psychological Measurement 53(3), 801-813 (1993)
15. Rietzschel, E. F., De Dreu, C. and Nijstad, B. A.: What are we talking about, when we talk about creativity? Group creativity as a multifaceted, multistage phenomenon. In: Mannix, E. A., Goncalo, J. A., Neale, M. A. (eds.), Creativity in Groups, Research on Managing Groups and Teams Series 12, 1-27. Emerald Group Publishing Ltd. (2009)
16. Rummery, G. A. and Niranjan, M.: On-line Q-learning using connectionist systems, Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University, UK (1994)
17. Russel, S. and Peter Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Inc. (2010)
18. Taggar, T.: Individual Creativity and Group Ability to Utilize Individual Creative Resources. The Academy of Management Journal 45 (2), 315-330 (2002)
19. Sternberg, R.J., Lubart, T.I. , Kaufman, J.C., Pretz, J. E.: Creativity. In K. J. Holyoak, K. J., Morrison, R. G. (eds.) The Cambridge handbook of thinking and reasoning, pp 351-369. New York: Cambridge University Press (2005)
20. Watkins, C. J. C. H.: Learning from Delayed Rewards. Ph.D. Thesis, University of Cambridge, UK, http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf, Accessed April 13, 2018
21. Yeh, Y. C.: The Effects of Contextual Characteristics on Team Creativity: Positive, Negative, or still Undecided. In: Roger Greatrex, Nina Brand (eds.). Working papers in Contemporary Asian Studies, 38, Centre for East and South-East Asian Studies, Lund University, http://lup.lub.lu.se/search/ws/files/3103258/3127683.pdf, Accessed April 13, 2018

Appendix A: *Gough Personality Scale*

Please indicate which of the following adjectives describe yourself the best. Check all that apply. The scoring key is between the brackets and it is not known by the people being evaluated.

_____ Capable (+)	_____ Honest (-)
_____ Artificial (-)	_____ Intelligent (+)
_____ Clever (+)	_____ Well-mannered (-)
_____ Cautious (-)	_____ Wide interests (+)
_____ Confident (+)	_____ Inventive (+)
_____ Egotistical (+)	_____ Original (+)
_____ Commonplace (-)	_____ Narrow interests (-)
_____ Humorous (+)	_____ Reflective (+)
_____ Conservative (-)	_____ Sincere (-)
_____ Individualistic (+)	_____ Resourceful (+)
_____ Conventional (-)	_____ Self-confident (+)
_____ Informal (+)	_____ Sexy (+)
_____ Dissatisfied (-)	_____ Submissive (-)
_____ Insightful (+)	_____ Snobbish (+)
_____ Suspicious (-)	_____ Unconventional (+)

Appendix B: *MSLQ (Motivated Strategies for Learning Questionnaire) adapted for Computer Science students*

Please rate the following items based on your beliefs on a 7 point scale where *1=not at all true for me* and *7= very true for me*. For anything in between choose a number between 2 and 6. Please rate each item choosing the scoring that suits you the best.

1.	Being enrolled in a Computer Science study program, I prefer classes that are challenging and trying to put me to test, so that I can learn new things.	
2.	Provided that I will study properly, then I will be capable of acquire the knowledge in the curricula.	
3.	When taking a test, I kepp thinking how low my results will be compared with other students.	
4.	I think that I will be capable to use what I have learned during university years to other training or study programs and jobs.	
5.	I think I will graduate with a good GPA.	
6.	I amm sure I can understand the most difficult materials or ideas thought or found in the obligatory readings.	
7.	Obtaining a good GPA at graduation is the most satisfying thing for me at this point.	
8.	When I am taking a test, I can not stop thinking about the parts that I cannot solve adequately.	
9.	It is my doing wrong if I will not be able to acquire the knowledge required as a	

	Computer Science graduate.	
10.	It is important to me to acquire the knowledge required as a Computer Science graduate.	
11.	Improving my GPA is the most important for me now, therefore my main concern is to get good grades.	
12.	I am confident that I can acquire the knowledge, abilities and skills required as a Computer Science graduate.	
13.	If I am able, I want to get better grades than most of my colleagues.	
14.	When taking tests or exams, I think of what may happen if I do not pass.	
15.	I am confident that I can understand the most comple material thought at this study program.	
16.	Being enrolled in a Computer Science study program, I prefer classes that make me curious even though they are difficult to understand.	
17.	I am very interested in the content of the courses thought at this study program.	
18.	If I try enough, I can understand the content of the courses thought at this study program.	
19.	I worry great deal about tests.	
20.	I believe that I can do an excellent job with regard to the given assignments, tests, and exams.	
21.	I expect I will be able to do well as a student of this study program.	
22.	The most satisfying thing for me as a student of this study program is to try understand the content of the courses as completely and as deeply possible.	
23.	I think that the instructional materials used for each course are useful and help me learn.	
24.	When given the opportunity during a class, I choose taks from which I can learn something new, even though that does not guarantee a good grade.	
25.	My not understanding of the content of the curricula thought is due to not working hard enough.	
26.	I like the subjects of the courses thought during this study program.	
27.	To understand the content thought is important to me.	
28.	I am very nervous when taking a test.	
29.	I am sure I can excell at the competencies achieved during this study program.	
30.	I want to do well during university years and at graduation because I want to show my capabilities to my family, friends, employeers or to others.	
31.	Taking into account the difficulty of this study program, the faculty and my abilities, I think I will do well as a student here.	

Towards Construction of Creative Collaborative Teams Using Multiagent Systems

Gabriela Moise *gmoise@upg-ploiesti.ro*
UPG University of Ploiesti
Ploiesti, Romania

Monica Vladoiu *monica@unde.ro*
UPG University of Ploiesti
Ploiesti, Romania

Zoran Constantinescu *zoran@unde.ro*
UPG University of Ploiesti
Ploiesti, Romania

Abstract

Group creativity and innovation are of chief importance for both collaborative learning and collaborative working, as increasing the efficiency and effectiveness of groups of individuals performing together specific activities to achieve common goals, in given contexts, is of crucial importance nowadays. Nevertheless, construction of "the most" creative and innovative groups given a cohort of people and a set of common goals and tasks to perform is challenging. We present here our method for semi-automatic construction of "the most" creative and innovative teams given a group of persons and a particular goal, which is based on unsupervised learning and it is supported by a multiagent system. Individual creativity and motivation are both factors influencing group creativity used in the experiments performed with our Computer Science students. However, the method is general and can be used for building the most creative and innovative groups in any collaborative situation.

Keywords: Creative Collaborative Working or Learning Groups, Multiagent System, Unsupervised Learning.

1. Introduction

Group creativity and innovation are of chief importance for both collaborative learning and collaborative working, as increasing the efficiency and effectiveness of groups of individuals performing together specific activities to achieve common goals, in given contexts, is of crucial importance nowadays. Therefore, educational institutions and companies alike have become more and more interested in increasing group creativity in both learning and working situations. *Creative learning* refers to instructional processes that have an extra focus on the development of creative abilities of individuals. *Collaborative creative learning* approaches creative learning that results from interactions and collaborations that take place between learners, while working together to fulfill common goals, and that has potential to enhance creativity both at individual level and group level. Moreover, collaborative creativity may be improved by providing appropriate environments and contexts and by organizing the individuals in suitable groups, as related work shows. However, it is still quite challenging to determine in which way the interactions and collaborations that take place inside a group result in either increases or decreases in creative group performances.

In this paper, we present a method of grouping individuals in *creative collaborative groups* whose creativity is increased iteratively. This method is based on an adapted version of the unsupervised learning algorithm introduced in [40]. The method has been introduced in

[19] and has been developed and evaluated further in [39], being under implementation with support from a multiagent system. This method and the corresponding architecture have been developed from scratch to help us in our continuous work of improving educational processes in which we are involved. The main contributions of the current work are the new architecture of the multiagent system, the algorithm for constructing and storing execution plans, the detailed presentation of an educational experiment performed with our Computer Science students, based on the proposed method, along with an updated and much more comprehensive overview of the related work.

However, the method is general and can be used for obtaining the most creative and innovative groups in any collaborative working or learning situation.

The structure of the paper is as follows: the next section includes the related work, the third one presents our multi-agent system for building creative groups that are involved in collaborative working or learning and with which we have done some preliminary tests in educational scenarios that are presented in Section 4, and the last section include some conclusions and future work ideas.

2. Related Work

In this section we overview the related work that includes three research directions, i.e. creativity in groups, modeling group creativity, and approaches similar to ours with regard to building creative groups. Creativity is a concept highly debated in psychological literature. Sternberg et al. view *creativity as the ability to produce work that is novel (i.e., original, unexpected), high in quality, and appropriate* [34]. Understanding creativity is challenging and has lead to elaboration of many theories, e.g. the *investment theory of creativity* [35, 36]. According to that, creative people are the ones *who are willing and able to, metaphorically, buy low and sell high in the realm of ideas*. Buying low means working on ideas that are not well-known or not popular that, however, have an intrinsic potential for growth. When introduced for the very first time, such ideas may face resistance, but creative people will fight it, and, in the end, they have an important opportunity to "sell" high, an innovative, influential, or popular idea, achieving this way a *creativity habit* [36]. Some authors point out that creativity is multifaceted and can be assessed by measuring *fluency* (creative production of nonredundant ideas, insights, problem solutions, or products)*, originality* (uncommonness or rarity of these outcomes)*, and flexibility* (how creativity expresses itself when using comprehensive cognitive categories and perspectives) [27].

Nevertheless, group creativity is a recent topic in the literature pointing to *the social nature of the creative act* [8]. Group creativity means more that summing up the individual creativities of the members, as the interactions that take place between them within the group, the diversity of members' backgrounds, abilities, and knowledge generate added value in creative processes. Thus, the importance of interactions between the group members and their role in stimulating creative processes contribute to increased group synergy. Several cognitive, social, and motivational factors influence the increase of group creativity such as: exchange of ideas, potential for competitiveness that allow individuals to compare their performances with the ones of their teammates, concept, product and perspective sharing, intrinsic motivation, openness to new experiences, etc. [3].

Amabile introduced the *componential theory of creativity*, along with the elements that influence creativity: at individual level (*domain-relevant skills, creativity-relevant processes, and task motivation*) and external (*the social environment in which the work takes place*). The domain-relevant skills refer to the knowledge and expertise of the individual in a specific field, while the creativity-relevant processes to individual characteristics that favor creativity: cognitive style, personality traits etc. Task-motivation is the internal individual motivation. Moreover, the author points out that *a central tenet of the componential theory is the intrinsic motivation principle of creativity* [2]. In his model of group creativity, Sawyer sees creativity as a synergy between *synchronic interactions* and *diachronic exchanges* [29]. While developing his *multilevel model of group creativity*, Taggar highlights that besides including creative members, team creativity is significantly influenced by *relevant processes that*

emerge as part of group interaction [38]. Moreover, creativity evolves over time within teams and is influenced by the *climate of creativity*, an essential feature in the *multilevel model of group creativity* of Pirolla-Merlo and Mann [25].

Contextual factors that influence creativity are divided in three categories [45]: (1) *facilitators of team creativity* (supervisory and co-workers support, psychological safety, group process), *(2) obstructers of generation of creative ideas* (conformity, insufficient resources, bureaucratic structure)*, and uncertain factors* (team diversity, conflicts in teams, group cohesion). *A*n *interactionist perspective on organizational creativity* is shown in the interactionist model of individual creative behavior of Woodman et al. Thus, group creativity is seen as *a function of individual creative behavior "inputs"*, *the interaction of the individuals involved* (e.g. group composition), *group characteristics* (e.g. norms, size, cohesiveness), *group processes* (e.g. approaches to problem solving), and *contextual influences* (e.g. the larger organization, the task). Moreover, organizational creativity is seen as *a function of the creative outputs of its constituent groups and contextual influences (organizational culture, reward systems, resource constraints, the larger environment, etc)*. This multifaceted mix boosts *the gestalt of creative output* (new products, services, ideas, procedures, processes, etc.). When building creative groups several characteristics may be considered, at various levels: *individual* (cognitive abilities/style, personality, intrinsic motivation, knowledge)*, group* (cohesiveness, size, diversity, role, task, problem-solving approaches)*, and organizational* (culture, structure, strategy, technology, resources, rewards etc.) [41], [43]. An outline for organization of group creative processes is proposed in [23]. A creative idea generation process was considered with respect to the social interactions inside the selected group, based on general principles from soft computing mathematical models.

Limited experiments with grouping individuals in creative groups are available in the literature. In [17], students involved in collaborative learning are grouped based on their learning styles. A research project that *investigates empirically whether knowledge sharing in community contexts can result in group knowledge that exceeds the individual knowledge of the group's members and concludes that this is the hallmark of collaborative learning* is available in [33]. An experimental study that worked on the assumption that shared cognition influences the effectiveness of collaborative learning and is crucial for cognitive construction and reconstruction of meaning is available in [37]. The work towards an intelligent collaborative learning system able to identify and target group interaction problem areas is available in [31]. Intense social interaction and collaboration are proven to provide for creation of learning communities that *foster higher order thinking through co-creation of knowledge processes* [15]. In [10], the "optimal class" is seen as *a high performing cooperative group with positive interdependence*. The issue of identifying peers and checking their suitability for collaboration, as an essential pre-collaboration task, is approached in [13], which concludes that a more personalized cooperation can take place provided that individual tastes and styles are considered. In [22], the authors approach the *liberating role of conflict in group creativity*, as a possible solution for weaknesses of group creativity, namely social loafing, production blocking, and evaluation apprehension. They have carried out an experiment in two countries to prove that brainstorming may benefit significantly from dissent, debate, and competing views, stimulating this way divergent and creative thought. In [26], the authors build up on two main ideas, namely that creative groups fuel both innovation and organizational change and that collaborative systems can be used to team up individuals across the globe in creative groups. They are concerned with the relation between individual creative preference and group creative performance across different phases of creative problem solving, in a group supported system. After experimenting with 250 students, their results indicate that *group member creative styles play an important role in determining the groups' productivity as well as certain qualities of the solution they pick*.

3. GC-MAS - A Multiagent System for Building Creative Teams

This section includes a brief presentation of our multi-agent system for building creative and innovative teams. The goal is grouping individuals in "the best" teams possible and our approach is innovative in the sense that grouping individuals in creative and innovative teams in an iterative semi-automated process has not been performed yet, up to our knowledge. This work builds up on previous work [19], where the very first architecture of the system was introduced. However, after experimenting with it, we have refined it further and reduced the number of agents, some of them having more complex roles, such as the facilitator agent.
The current system architecture includes the following agents, in which all the agents are task agents, except for CommGC (Fig. 1):

- *The Communication Agent (CommGC)* has a dual role, being responsible with interfacing with the users (both students and instructors) and with the agents, along with managing the activities of the other agents;
- *The Creative Groups' Builder (BuildGC)* is an agent that assists the construction of creative groups based on an unsupervised learning algorithm;
- *The Creativity Evaluation Agent (EvalGC)* assesses each group creativity;
- *The Creativity Booster (EnvrGC)* boosts development and maintenance of contextual environments that provide for increasing group creativity;
- *The Facilitator Agent (FclGC)* facilitates a more efficient group interaction, e.g. by sustaining the team members who are shyer or less active. It also provides support for seeking out and taking on otherwise neglected tasks that have potential to facilitate creative group performances.

CommGC acts as a middle agent and has a horizontally stratified structure, in which each level is connected directly to both the input sensors and the output effectors (software entities that perform particular actions). Each level acts as an individual agent that provides the expected action. The two levels of CommGC are as follows: (1) *the social level* that ensures the communication with the other agents, the users, and with the external environment, as a true personal/interface agent, and (2) *the administrative level* that coordinates the actions of all the agents (see Fig. 2).

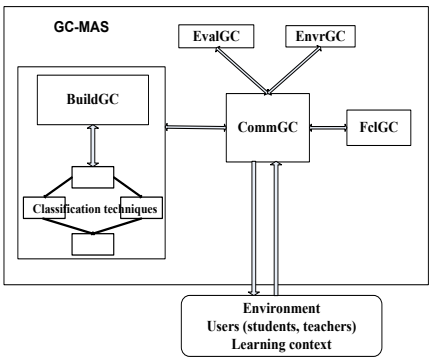

Fig. 1. GC-MAS - the bird's eye view architecture.

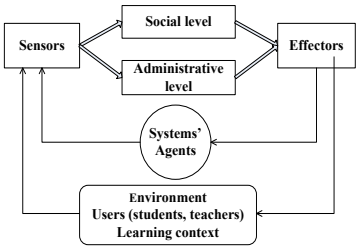

Fig. 2. The architecture of CommGC.

The agents *BuildGC, EvalGC, EnvrGC* and *FclGC* are execution agents that perform precise actions in construction of creative groups. They have a very simple structure, are goal-oriented, and use plan libraries or classification techniques to perform their duties, as it can be seen in Fig. 3. At the core of execution agents is their plan library, as *planning is essentially automatic programming: the design of a detailed course of action which, when executed, will result in the achievement of some desired goal* [44]. *A plan library (PL)* is defined by a set of inputs (plans) $PL=\{P_1, P_2, ..., P_n\}$, which an agent uses to achieve its goals. Such an input includes the plan's pre-conditions, body, and its post-conditions. A plan P_i is defined as $P_i=<pre_i, body_i, post_i>$. *The pre-condi*tion is defined by a logical expression and each time the value of this expression is true the specified/associated plan is executed. *The post-*

condition specifies the goal that an agent is supposed to fulfill. *The body* of a plan is a computer program specified by a sequence of primitive actions that is executed when its pre-condition is true (1).

$$< actions\_sequence >=< primitive\_action >< actions\_sequence > |NULL \quad (1)$$

The plans are built using a constructor. One of the most well-known algorithms for this purpose is the STRIPS planning algorithm, in which a means-ends analysis is performed to find an action sequence that will lead to achieving the goal [6]. Planning is seen as a search of an action sequence in a state space based on the pre-conditions and on the outcomes of the actions. Another approach consists in adaptation of the existing plans to a specific situation (case based reasoning) [1]. The plan constructor is seen as a black box that returns a plan solution given a plan description. In GC-MAS, we use the algorithm for constructing and storing a plan in Fig. 4. First, we abstract the state of the system and its goal and we model them with a conjunction of primitive states (2), respectively of primitive goals (3) i.e. that cannot be decomposed any further. For example, primitive states could be *the learning style is visual or the motivation of the student is intrinsic*. A primitive rule is defined as follows: *if state then primitive_action*. A priority function is associated to each primitive rule P: R→N, where R is a set of rules and N is the set of natural numbers. The priority function helps solving the selection conflict when for the same pre-condition more than one action may be chosen. In such cases, the action with the highest value of priority function will be selected. The primitive actions and rules are stored in libraries available to each agent. The algorithm generates a plan that leads the system to achieve the goal **g** starting from a state **st**. Two situations are similar if their composing states and goals are similar. Two states *State1* and *State2*, respectively two goals *Goal1* and *Goal2* are similar if their similarity index is above a fixed threshold (4, 5).

$$State = st_1 \wedge st_2 \wedge \ldots \wedge st_n \quad (2)$$

$$Goal = g_1 \wedge g_2 \wedge \ldots \wedge g_m \quad (3)$$

$$State1 = st_{11} \wedge st_{12} \wedge \ldots \wedge st_{1n}\ State2 = st_{21} \wedge st_{22} \wedge \ldots \wedge st_{2m} \quad (4)$$

$$state\_index_{similarity} = |\{st_{11}, st_{12}, \ldots, st_{1n}\} \cap \{st_{21}, st_{22}, \ldots, st_{2m}\}|$$

$$Goal1 = g_{11} \wedge g_{12} \wedge \ldots \wedge g_{1n}\ Goal2 = g_{21} \wedge g_{22} \wedge \ldots \wedge g_{2m} \quad (5)$$

$$goal\_index_{similarity} = |\{g_{11}, g_{12}, \ldots, g_{1n}\} \cap \{g_{21}, g_{22}, \ldots, g_{2m}\}|$$

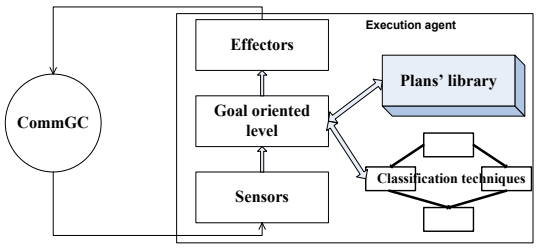

Fig. 3. The architecture of an execution agent.

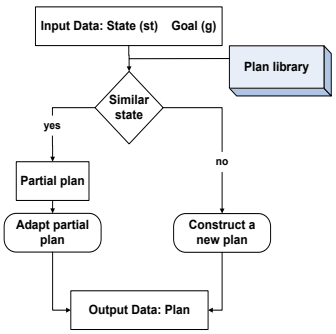

Fig. 4. The algorithm for plan construction.

Case I. A similar situation does exist, so there is a plan whose pre-condition is similar with the system state and the plan post-condition is similar with the desired goal (Fig. 5). This plan is selected, adapted if necessary for the similar situation, and then stored in the plan library. The procedure for plan adaptation is as follows:

- If the system state contains the plan pre-condition and the agent's goal is included in the plan post-condition then the plan remains unchanged;
- If a goal that is not included in the plan post-condition exists then a backward search is performed in the state space (built from the plan libraries and rules) to

determine a sequence of primitive actions that leads to that goal, given the system's state. This particular sequence of primitive actions is included in the selected plan to obtain its adaptation to a similar situation.

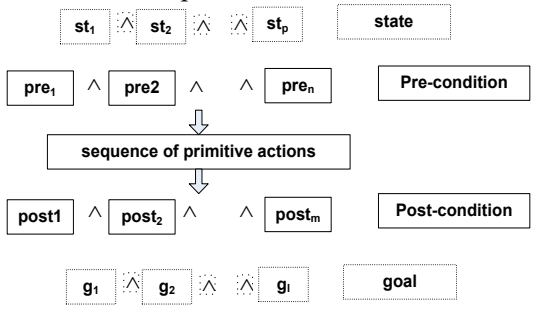

Fig. 5. A similar situation exists.

Case II. A similar situation does not exist

For each sub-goal g_i of the goal, a sequence of primitive actions is searched so that their execution leads to the desired goal starting from a particular state. The action sequences that are found this way are further combined to form the body of a plan.

BuildGC - The Creative Groups' Builder aims at construction and iterative refinement of creative groups taking into account factors that boost creativity, their interdependencies and the purpose of building of particular creative groups. *The input data* for BuildGC are *student data* (individual features that influence group creativity), *group data* (the purpose of constructing creative groups, i.e. the problem to be solved, the task to be completed, the research to be undertaken etc., the group size, the diversity of group members, etc., and support data generated by both users and other agents autonomously or as a result to the queries addressed by *BuildGC*. *The output data of BuildGC* consists of both the most creative learning groups buildable and the queries to other users and agents with respect to the process of group construction. In our experiments, BuildGC had the plan structure as follows: the pre-conditions consisted of each student's creativity features, the body consisted in a prediction reasoning tool based on an adapted version of the Q-learning algorithm [19], [40], while the post-condition included the best organization of a cohort of students in creative groups so that the value of Q is the largest possible for each group. In brief, this algorithm is a reward learning algorithm that starts with an initial estimate *Q(s, a)* for each pair *<state, action>*. When a certain action a is chosen in a state s, the intelligent system (the agent *BuildGC* in our case) gets a *reward* R(s, a) and the next state of the system is acknowledged. The function value-state-action is estimated as:

$$Q(s,a) \coloneqq Q(s,a) + \alpha\big(R(s,a) + \gamma max_{a'} Q(s',a') - Q(s,a)\big) \quad (6)$$

Where $\alpha \in (0,1)$ is the learning rate, $\gamma \in (0,1)$ is the discount factor, and s' is the state reached after executing the action a in the state s. The way in which the values for the learning rate and for the discount factor should be selected is discussed in [14]. Value 0 for the learning rate means that the value for Q is never updated and that the system never learns. Selection of a higher value means that learning is faster. In our first experiments, we used a 0.5 learning rate. The discount factor has values between 0 and 1. Closeness to 1 means that a future reward is more important to the system than an immediate reward, i.e. that the importance of a future reward is increased, as γ is still below 1. A balance between the immediate rewards and the past rewards is sought for in dynamic environments.

From GC-MAS's point of view, the environment consists in the students, the instructor, and the learning context (as in [28]). For *BuildGC*, the agent that computes the best grouping of a cohort of students in creative teams, the environment is the structural organization in a set of groups. However, the groups' structure changes over time, as the agent learns from its interactions with its environment how to construct more and more creative groups. To fulfill its goal of building the most creative groups, *BuildGC* uses the GC-Q-learning adapted algorithm [19]. In this case, the reward is the "value of group creativity" that ranges between

1 and 5. The goal here is to obtain a final state, *namely an optimal organization of students in groups*, in which either each group will have a creativity value larger than a desired threshold or the average creativity on all the groups will be higher than such a threshold. The GC-Q-learning algorithm is as follows:

1. Build a bi-dimensional matrix *Q* for all the possible pairs <*state, action*>. The columns of this matrix consist of *(c_1, c_2, ..., c_m, no_group, action_number, q)*. A value of the *action_number* of *i* means that if a particular type of student (given by his creativity vector c_1, c_2, ..., c_m) will be moved to the group having the value of *no_group i* then her contribution to group creativity is quantified by q (in this stage). All the elements in the q column may be initialized with 0 or with a randomly chosen low value. On each line of the matrix, the data that corresponds to each type of student involved in the grouping process is included, i.e. the values of his characteristics, the current group number, the action number, and the value computed for *q (that quantifies a potential for creativity)*. One particular type of student could have more related lines, one for each combination <*current group number, action*>;
2. Initialize the *optim_policy* with an initial policy. In our case, the optimal policy is the optimal grouping of students for boosting group creativity. The initial grouping is set by the instructor and the students together and generally they group as cliques;
3. Group the students and have them carry on working sessions, in which each group's creativity is assessed and its score is assigned to the reward *R(s,a)*. The values of *R(s,a)* are obtained for now with help from human experts. We may say that R *materializes* that potential for creativity (q). Then, the matrix *Q* is re-calculated for each such working session. This procedure is shown below.

```
procedure working_session_computation
select action of (optimal_policy)  /* student grouping*/
compute R(s,a)
compute table Q /* using formula (6)*/
```

4. Analyze the group creativity for each group against the global objective (the optimal grouping policy), which is getting closer to the maximum value possible for R, for each group or for all the groups. Re-iterate from step 3, if necessary.

Once the optimal policy consisting in tuples *(c_1, c_2, ..., c_m, group number)* is obtained and *BuildGC* has learned enough, *predictions may be made for each new type of student, given his set of characteristics*. The predictions consist of a series of group numbers, which are presented sorted decreasingly according to the contribution made by that particular generic student to each group's creativity. Thus, the first number in the series is of the group in which that generic student would contribute the most to the group creativity, the second one of the group in which she would make the second best contribution, and so on. Other classification techniques may be used as well (neural network based classifiers, Bayes classifiers, decision trees, or support vector machines). A detailed description of the Bayesian networks-based classification techniques can be found in [7], [11]. We have already worked on this idea of building the most creative and innovative collaborative groups using Bayes classifiers with encouraging results [18].

EvalGC - The Creativity Evaluation Agent supports assessing of group creativity based on criteria for measuring ideation, namely novelty, variety, quantity, and quality [30]. It uses a plan library to achieve its goals of (1) recording the ideas generated by the group and classifying them, (2) calculating the frequency of good ideas' production (as the number of innovative and useful ideas per time unit), and (3) keeping the creativity score and ensuring the communication via *CommGC*.

EnvrGC - The Creativity Booster aims to enhance group creativity by providing for contextual environments that include consistent activators that contribute to creativity boosting. The agent works by "pushing on" the creativity triggers specific to the situation. In

our case, this action can be performed using a fuzzy controller with which we have worked previously [20].

Facilitator Agent-FclGC provides for a more efficient group interaction, e.g., by sustaining the team members who are shyer or less active, and by supporting seeking out and taking on otherwise neglected tasks that have potential to increase creative group performances. The execution plans of this agent are presented below:

FclGC - Execution plan 1
Pre-condition: whenever the number of ideas generated per minute is more than 10;
Body: the agent asks the online group members to focus on the task to do, following their common goal; specific creativity triggers: advising; motivation;
Post-condition: group refocuses on the task at hand, draws some conclusions.

FclGC - Execution plan 2
Pre-condition: whenever a group member has not been active, generating ideas or contributing to the discussions for 5 minutes;
Body: the agent asks that member to say a new idea or to make a comment on what it has been said so far; specific creativity triggers: advising; motivation;
Post-condition: a new idea/comment made by the less active member is generated.

FclGC pro-actively prevents situations in which group members focus entirely on coming up with their own ideas and ignore completely (to build on) the ideas of others, which is an essential added value of working together in a group [4]. For this situation, the execution plan of FclGC is as follows:

FclGC - Execution plan 3
Pre-condition: every 15 minutes or every 25 ideas generated;
Body: the agent asks the online group members what they think about the ideas generated so far and if they could build up on them for a while instead of generating new ideas; specific creativity triggers: reviewing and replaying session histories;
Post-condition: students overview previous ideas and build up on them for 5 minutes.

4. A Real World Educational Experiment

To use this method, one needs to initially group the students randomly or based on their interpersonal affinities, then have them work as groups in a particular (educational or working) scenario, after which their group creativity can be assessed. Based on their creativity characteristics and using the adapted Q-learning algorithm, the composition of the groups may change in order to reach the global creativity objective. The goal here is to obtain a final state, *namely an organization of students in groups*, in which either each group will have a creativity value larger than a desired threshold or the average creativity on all the groups will be higher than such a threshold). Further on, the obtained data (group creativity is the reward of the algorithm) is fed back to the algorithm and, this way, *it learns over time what is the best option of moving a (particular type of) student in the group in which s/he has the maximum contribution to the group's creativity*. Globally, for a pool of students, the objective is to group the students so that the global creativity goal is reached [39].

After clarifying the conceptual aspects of GC-MAS, we have been concerned with investigating the viability of our approach and therefore we have tested it in some educational scenarios with our Computer Science students (both undergraduate and graduate). In this section, we present briefly an educational experiment performed using the proposed approach. More details about a similar larger experiment may be found in [39]. The main stages of the experiment have been as follows:

1. *The evaluation of each student's individual creativity and motivation using several evaluation tools*. To assess individual creativity, we have used both the Gough Creative Personality Scale [9] [39] and an extended version of the Creative Achievement Questionnaire [4] that we have adapted for Computer Science students.

We present here the data obtained using Gough Scale, which is simpler and easier to understand. Generally, the Gough Score values range between -12 and 18. The student motivation can be low (having value 0), middle (1), or high (2) and it has been determined using our adapted questionnaire based on MSLQ (Motivated Strategies for Learning Questionnaire) [24] [39].

2. *Initial organization of students in groups based on their inter-personal affinities.* Have them carry the first online brainstorming session. Evaluation of the group creativity for each group. If the global objective has been reached then stop.
3. *Activation of the BuildGC agent for the students' cohort to group them in the most creative groups possible.* First, this agent will indicate for each student to which group will contribute the most to group creativity. Based on that, a student may be moved to a group for which his q value is among first 30% in decreasing order (*to raise the potential for increasing group creativity*). Then the collaborative creative activity takes place, in our case a second online brainstorming session.
4. *Evaluation of group creativity for each group involved in the experiment*. If the global creativity objective has not been reached, re-iterate from stage 3.

The experiment included three online brainstorming sessions on subjects of interest for them: (1) the improvement of both the curricula and the syllabuses for our Computer Science programs (undergraduate and graduate), (2) the preferred teaching and learning methods, and (3) the enhancement of their student life within university and campus. Each session had to end with a final conclusion on the issues discussed. We used brainstorming here just for measuring group creativity, but any kind of appropriate evaluation can be used.

For this experiment, the *Q* matrix had 45 lines and 5 columns. Each column consists in, respectively, the Gough score, the motivation value, the current group number, the action number (that means to move her in the group in which she would contribute the most to group creativity, given her characteristics), and the *q* value. On each line of the matrix we have the data that correspond to each type of student involved in the grouping process, i. e. the values for: the Gough score, the motivation, the current group number, the action number, and the value of *q*. We present below some experimental results obtained while trying to group in increasingly creative teams several pools of students having various values for *the creativity pair* (Gough score, motivation value). In this experiment, we had 5 types of students characteristic-wise with these pairs as follows: (3,1), (3,2), (2,1), (2,2), and (4,1), and we have studied 9 possible groups. In Table 1 the sample data for the students having the pairs (2,1) and (4,1) are shown. The interpretation of this data is that a student with the pair (2,1) would contribute the most to the group creativity if s/he would be in group 2, and decreasingly - in group 5, 7, 8 or 4. A student with (4,1) would contribute the most to the group creativity if s/he would be in group 3, and decreasingly - in group 5, 7, 9, or 6.

Table 1. Sample Data for Students with Creativity Pair (2,1) – left and (4,1) – right.

Gough score	Student motivation	Action – move to group no	Q value	Gough score	Student motivation	Action – move to group no	Q value
2	1	1	0	4	1	1	0
2	1	2	**3,5**	4	1	2	0
2	1	3	0	4	1	3	**3,78875**
2	1	4	1,9	4	1	4	0
2	1	5	2,705	4	1	5	2,777188
2	1	6	0	4	1	6	2,277188
2	1	7	2,54	4	1	7	2,612188
2	1	8	2,54	4	1	8	0
2	1	9	0	4	1	9	2,612188

However, the individuals are not grouped and re-grouped indefinitely, as the algorithm learns during time in which group a person should be to contribute the most to group's creativity. So, it can make a recommendation in this sense. In our particular case, during our work with the students involved, throughout their university years, both as undergraduate and graduate, we have evaluated the creativity of the teams obtained in this way and the results show that they are, indeed, more creative than ad-hoc or buddy teams, as they consistently obtain better evaluations of teamwork results [18, 19], [39]. But the method is general and can be used in any collaborative working situation where increasing group creativity is required.

5. Conclusions and Future Work

We introduced here our semi-automated method of grouping team members in increasingly creative groups, which has been tested using a multiagent system prototype. Moreover, we have performed some experiments, the results being encouraging so far. Thus, our first results show that students can be more creative provided that they are included in appropriate groups for activities that involve teamwork [18] [19] [39]. The importance of taking into account how teams are made for such activities is pointed out once again in accordance with the results of other similar research [10], [12], [13], [15], [17], [22], [31], [33], [37]. It seems to make more sense to apply this semi-automatic grouping method for groups of people aiming at becoming teams, over long periods of time, such as university or working years. Though, the method can be used also for groups formed for shorter durations because it is based on features that quite often have the same values for different people (for instance, the creativity pair <individual creativity, motivation>), so the process does not need to start from scratch each time, but just build up on previous results. More tests on various scenarios need to be performed, in various learning or working activities, with diverse pools of individuals, using control groups, and so on. More factors that influence group creativity need to be taken into account too, for example, group interactions and the way they develop over time.

Development of a software tool that implements the method presented here would be very useful to assist in construction of the most creative and innovative groups in particular learning or working contexts and in other collaborative scenarios as well. Other future work ideas include corroborating the results obtained with several creativity evaluation scales, using metrics to evaluate group creativity, inclusion of contextual and organizational factors, improving the algorithm, and, finally, offering the method as an online open service.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications. 7(1), pp. 39-59. IOS Press (1994).
2. Amabile, T. M.: Componential Theory of Creativity. In: Kessler, E. H. (ed.), Encyclopedia of Management Theory, pp. 135-140, SAGE Publications Inc. (2013).
3. Baruah, J., Paulus, P. B.: Enhancing Group Creativity: The Search for Synergy. In: Mannix, E. A., Goncalo, J. A., Neale M. A. (eds.), Creativity in Groups , Research on Managing Groups and Teams Series. 12, pp. 29-56, Emerald Group (2009).
4. Bolinger, A. R., Bonner, B. L., Okhuysen, G. A.: Sticking together: the glue role and group creativity. In: Mannix, E. A., Goncalo, J. A., Neale M. A. (eds.), Creativity in Groups, Research on Managing Groups and Teams Series. 12, pp. 267-291, Emerald Group (2009).
5. Carson, S., Peterson, J. B., Higgins, D. M.: Reliability, Validity, and Factor Structure of the Creative Achievement Questionnaire. Creativity Research Journal. 17(1), pp. 37–50. Taylor & Francis (2005).
6. Fikes, R. E., and Nilsson, N.: STRIPS: A new approach to the application of theorem proving to problem solving. Artificial Intelligence. 5(2), pp. 189-208. North-Holland Publishing Company (1971).

7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning. 29, pp. 131–163. Kluwer Academic Publishers (1997).
8. Gorny, E. (ed.): Group creativity. Dictionary of Creativity: Terms, Concepts, Theories & Findings in Creativity Research, (2007), http://creativity.netslova.ru/Group_creativity.html. Accessed April 17, 2018.
9. Gough, H. G.: A Creative Personality Scale for the Adjective Check List. Journal of Personality and Social Psychology. 37, pp. 1398-1405. American Psychological Association (1979).
10. Israel, J., Aiken, R.: Supporting Collaborative Learning with an Intelligent Web-based System. International Journal of Artificial Intelligence and Education. 17(1), pp. 3-40. Springer (2007).
11. Joyce, J.: Bayes' Theorem. In: Zalta E. N. (ed.), The Stanford Encyclopedia of Philosophy (2003) http://plato.stanford.edu/entries/bayes-theorem/. Accessed April 17, 2018.
12. Koschmann, T.: Dewey's Contribution to the Foundations of CSCL Research. In: Proceedings of Conference on Computer Support for Collaborative Learning, Foundations for a CSCL Community, pp. 17-22, Boulder, CO, USA, (2002).
13. Kumar, V.: Computer Supported Collaborative Learning - Issues for Research, https://pdfs.semanticscholar.org/2ecb/60766155e3bc0e435ae63964b93148b2adfa.pdf. Accessed April 17, 2018.
14. Leon, F., Şova, I., Gâlea, D.: Reinforcement Learning Strategies for Intelligent Agents. In: Proceedings of the 8th International Symposium on Automatic Control and Computer Science, Iaşi (2004).
15. Ada W. W. Ma: Computer Supported Collaborative Learning and Higher Order Thinking Skills: A Case Study of Textile Studies. The Interdisciplinary Journal of E-Learning and Learning Objects. 5, pp. 145-167 (2009).
16. Mannix, E., Goncalo, J. A., Neale, M. A.: Creativity in Groups, Research on Managing Groups and Teams Series, 12. Emerald Group Publishing Limited (2009).
17. Martin, E., Paredes, P.: Using Learning Styles for Dynamic Group Formation in Adaptive Collaborative Hypermedia Systems. In: Proceedings of the 1st International Workshop on Adaptive Hypermedia and Collaborative Web-based Systems, pp. 188-198 (2004).
18. Moise G., Vladoiu M., Constantinescu Z.: Building the Most Creative and Innovative Collaborative Groups Using Bayes Classifiers. In: Panetto H. et al. (eds) On the Move to Meaningful Internet Systems. OTM 2017 Conferences. Lecture Notes in Computer Science, 10573, pp. 271-283. Springer, Cham (2017).
19. Moise, G., Vladoiu, M., Constantinescu, Z.: GC-MAS - a Multiagent System for Building Creative Groups used in Computer Supported Collaborative Learning. In: Proceedings of the 8th International KES Conference on Agents and Multi-agent Systems – Technologies and Applications, Advances in Intelligent Systems and Computing. 296, pp. 313-323. Springer, Cham (2014).
20. Moise, G.: Fuzzy Enhancement of Creativity. In: Chiu D. K.W. et al. (eds.), New Horizons in Web Based Learning, LNCS, 7697, pp. 290-299. Springer-Verlag (2014).
21. Moise, G.: Triggers for Creativity in CSCL. In: Proceedings of the 9th International Scientific Conference eLearning and Software for Education, pp. 326-331, Editura Universitatii Nationale de Aparare "Carol I", Bucuresti (2013).
22. Nemeth, C. J., Personnaz, B., Personnaz, M., Goncalo, J.A.: The Liberating Role of Conflict in Group Creativity: A Study in Two Countries. European Journal of Social Psychology. 34 (4), pp. 365–374. John Wiley & Sons Ltd (2004).
23. Petry, F. E., Yager, R. R.: Principles for organization of creative groups. Journal of Ambient Intelligence and Humanized Computing. 5(6), pp. 789-797. Springer International Publishing (2014).
24. Pintrich, P. R., Smith, D. A., Garcia, T., Mckeachie, W. J.: Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educational and Psychological Measurement. 53(3), pp. 801-813. Sage (1993).

25. Pirola-Merlo, A., Mann, L.: The Relationship Between Individual Creativity and Team Creativity: Aggregating Across People and Time. Journal of Organizational Behavior. 25(2), pp. 235–257. John Wiley & Sons Ltd (2004).
26. Ray, D. K., Romano Jr., N. C.: Creative Problem Solving in GSS Groups: Do Creative Styles Matter? Group Decision and Negotiation. 22(6), pp. 1129-1157. Springer Science+Business Media B.V. (2013).
27. Rietzschel, E. F., De Dreu, C. K. W., Nijstad, B.A. : What are we talking about, when we talk about creativity? Group creativity as a multifaceted, multistage phenomenon. In: Mannix, E. A., Goncalo, J. A., Neale M. A. (eds.), Creativity in Groups, Research on Managing Groups and Teams Series. 12, pp. 1-27, Emerald Group (2009).
28. Russel, S. and Peter Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Inc. (2010).
29. Sawyer, R. K.: Group creativity: Music, theater, collaboration. Psychology Press (2003).
30. Shah, J. J., Vargas-Hernandez, N.: Metrics for measuring ideation effectiveness. Design Studies. 24(2), pp. 111-134. Elsevier Science (2003).
31. Soller, A., Abu Issa, A. S.: Supporting Social Interaction in an Intelligent Collaborative Learning System. Internationa Journal Artificial Intelligence in Education. 12(1), pp. 40-62. Springer (2001).
32. Stahl, G., Koschmann, T., Suthers, D.: Computer-Supported Collaborative learning: An historical perspective. In: Cambridge Handbook of the Learning Sciences, Sawyer R. K. (ed.), pp. 409-426, Cambridge University Press, Cambridge, UK (2006).
33. Stahl, G.: Cognition in Computer Assisted Collaborative Learning. Journal of Computer Assisted Learning. 21, pp. 9-90. Blackwell Publishing Ltd. (2005).
34. Sternberg, R. J., Lubart, T. I., Kaufman, J. C., Pretz, J. E.: Creativity. In: The Cambridge Handbook of Thinking and Reasoning, K.J. Holyoak, R.G. Morrison, pp. 351-369, Cambridge University Press, New York (2005).
35. Sternberg, R. J., Lubart, T. I.: An Investment Theory of Creativity and its Development. Human Development. 34(1), pp. 1–31 Karger International (1991).
36. Sternberg, R. J.: The Assessment of Creativity: An Investment-Based Approach. Creativity Research Journal. 24(1), pp. 3–12. Taylor & Francis (2012).
37. Stoyanova, N., Kommers, P.: Concept Mapping as a medium of shared cognition in Computer-Supported Collaborative Learning. Journal of Interactive Learning Research. 13(1), pp. 111-133. AACE (2012).
38. Taggar, S.: Individual Creativity and Group Ability to Utilize Individual Creative Resources. The Academy of Management Journal 45(2), pp. 315-330. (2002).
39. Vladoiu M., Moise G., Constantinescu Z.: Towards Building Creative Collaborative Learning Groups Using Reinforcement Learning, accepted for publication at the 27th International Conference on Information Systems Development ISD 2018 (2018).
40. Watkins, C.: Learning from Delayed Rewards, PhD Thesis, pp. 95-96, University of Cambridge, England (1989).
41. Woodman, R. W., Sawyer, J. E., Griffin, R. W.: Toward a Theory of Organizational Creativity. Academy of Management Review. 18(2), pp. 293-321 (1993).
42. Woodman, R. W., Schoenfeldt, L. F.: An Interactionist Model of Creative Behaviour. Journal of Creative Behavior. 24(4), pp. 279-290. John Wiley & Sons Ltd (1990).
43. Woodman, R. W., Schoenfeldt, L. F.: Individual Differences in Creativity: An Interactionist Perspective. In: Glover, J. A., Ronning, R. R., Reynolds, C. R. (eds.), Handbook of Creativity, pp. 77-92, Plenum Press, New York (1989).
44. Woolbridge, M. and Jennings, N. R.: Agent Theories, Architectures, and Languages: a Survey. In: Intelligent Agents, LNCS 890, pp. 1–22, Springer-Verlag (1995).
45. Yeh, Y. C.: The Effects of Contextual Characteristics on Team Creativity: Positive, Negative, or still Undecided. Working papers in Contemporary Asian Studies, 38, Lund, Sweden: Centre for East and South-East Asian Studies, Lund University, http://www.lunduniversity.lu.se/lup/publication/3127670. Accessed April 17, 2018.

Building the Most Creative and Innovative Collaborative Groups Using Bayes Classifiers

Gabriela Moise[1], Monica Vladoiu[1], Zoran Constantinescu[1]

[1] CerTIMF Research Center, Petroleum-Gas University of Ploiesti, ROMANIA
{gmoise,mvladoiu,zoran}@upg-ploiesti.ro

Abstract. Building "the best" creative and innovative groups that have common goals and tasks to perform, efficiently and effectively, is difficult. The complexity of this undertaking is significantly increased by the necessity to first understand and then measure what "the best" goal means for the individuals in the groups, but also for each group as a whole. We present here our Bayes classifiers-based technique for building "the best" groups of students to work together in collaborative learning situations. In our case, "the best" goal means *the most creative and innovative teams* possible in a given learning situation based on some particular attributes: *individual creativity*, *motivation*, *domain knowledge*, and *inter-personal affinities*. However, both the proposed model and method are general and they may be used for building collaborative groups in any situation, with the appropriate "the best" goal and attributes. A case study on using this method with our Computer Science students is also included.

Keywords: Creative and Innovative Collaborative Learning, Collaborative Work, Bayes Classification.

1 Introduction

Construction of groups that have common goals and tasks to perform in collaborative situations, e.g. in working or learning scenarios, is very usual nowadays. During cooperative activities, individuals work together to accomplish shared objectives and to obtain results beneficial to both themselves and other group members. However, coming up with a technique to build "the best" groups of people to optimally collaborate, both efficiently and effectively, while pursuing the achievement of common aims is not straightforward. The complexity of this undertaking is significantly increased by the necessity to first understand and then measure what "the best" means for the individuals in the groups, but also for each group as a whole. Therefore, in any given collaborative context, clustering the most suitable individuals to form the highest performance groups aiming at reaching some specific outcomes is very challenging.

Cooperative learning consists of the instructional use of small groups of students that work together to maximize their own and each other's learning. Thus, after receiving instruction from their instructor, class members are usually split into small groups that work through the assignment until all the group members have successfully understood and completed it [1]. But ad hoc grouping of students does not neces-

sarily mean that cooperative groups have been built, given that the purpose of cooperation is to maximize both achievement of the proposed goals and learning. Based on the performance in this respect, several categories of learning groups exist, namely pseudo, traditional classroom, cooperative, and high-performance cooperative learning groups [2]. For example, in cooperative learning groups students work together to accomplish common aims, while seeking outcomes beneficial to all group members. Students embark together on the learning journey, helping each other throughout the way. The main benefit is that the group becomes more than the sum of its parts and that virtually all students perform higher than they would if they worked alone. The high-performance ones outperform all the expectations for cooperative learning groups, the group members being highly committed to each other and to the group's success, so that the group becomes a team. The benefits of cooperative learning in educational processes are plentiful [1-4]. Thus, students learn more by doing and by involving actively and cooperatively in learning activities. Less motivated students can be appealed to continue working on difficult aspects by their teammates. Even strong students can benefit from clarifying the approached issues to their colleagues by improving their own understanding and mastering. Moreover, timely delivery of assignments is much more frequent [3].

We present here our work towards developing a technique, based on Bayes classifiers, for building "the best" groups of students to work together in collaborative learning situations. In our case, "the best" goal means *the most creative and innovative teams* possible, in a given learning situation, based on some particular attributes: *individual creativity*, *motivation*, *domain knowledge*, and *inter-personal affinities*. Our focus is dual: first, we introduce both a model and a method for grouping individuals in creative groups in collaborative situations (using Bayesian Networks-based classification), and, second, we instantiate and apply them in learning contexts. However, *both the proposed model and method are general* and they may be used for building collaborative groups in any situation, with the appropriate goal and attributes for that context. In the long run, we aim at the development of an intelligent system able to support organizing individuals in the most creative and innovative groups, in any given collaborative situation, and at offering it as an open project [5]. A case study on using this method with our Computer Science students is also included.

The structure of the paper is as follows: the next section presents the related work; the third one introduces both our model and method for building collaborative groups with instantiation for learning situations. Section 4 presents the results of our experimenting with the resulted technique, while in Section 5 they are evaluated and discussed. The last section includes some conclusions and future work ideas.

2 Related Work on Collaborative Creativity

The approach of creativity in the literature has shifted from focusing on gifted individuals to acknowledging that each person can be creative, and, finally, to recognizing that social structures have a strong influence on individual and group creativity [6-10]. Moreover, research shows that in order to understand how an individual contrib-

utes to group creativity a large variety of factors needs to be considered (individual characteristics, organizational environment, social relationships, etc.) as such and also combined. In [6], three research directions are proposed: group creativity in context, group-level creative synergy, and strategies for developing group creativity. In [7], the authors approach the factors that influence team creativity and innovation based on the triad Input–Process–Output. The Input shows the team composition based on the members' individual characteristics. The Process includes the activities undertaken by the team members to carry out some tasks or to solve some problems, while the Output consists of the creativity and innovation of the team (team effectiveness).

One of the most representative models for collaborative creativity is introduced in [8]. The input variables for this model are Group Member Variables, Group Structure, Group Climate, and External Demands. Three categories of processes are taken into account: cognitive, motivational, and social. The output consists of team creativity and innovation. Using this model, the authors show how the group member attributes (personality, task relevant knowledge, skills, and abilities, intrinsic motivation, cognitive flexibility, creative self-efficacy, etc.), the group structure (diversity, size, communication mode, cohesiveness, leadership style etc.), the group climate (commitment to task, conflict, trust, norms of participation/risk-taking/innovation etc.), and the external demands (creative mentors and models, rewards and penalties, freedom/autonomy/self-management, support for creativity, intergroup and intra-group competition, task structure, performance feedback etc.) influence the cognitive, social, and motivational processes that collaborative creativity relies on.

Despite the interest on increasing group creativity, a few experiments of grouping people in the most creative groups exist, in general, and, in learning, in particular. Moreover, most of them do not use data mining techniques, machine learning, nor intelligent data analysis. Thus, in [4], it is shown that in case of one wanting to teach her course effectively, ability heterogeneity should be her primary criterion. Also, if the groups need to meet outside class, forming teams of students who have common blocks of unscheduled time could be suitable. This work also points out some of the downsides of groups composed exclusively of strong students, who are likely to distribute the work rather than engaging in the group discussions and informal tutoring sessions that lead to many of the confirmed instructional benefits of cooperative learning. In [11], the authors have analyzed the cause-effect relationships between 6 factors: team creativity, exploitation, exploration, organizational learning culture, knowledge sharing, and expertise heterogeneity. They have also built a General Bayesian Network, which has as a target node the team creativity and that shows the dependencies between these factors. The main question addressed in this work was dual, i.e. (1) how do the processes of creative revelation—exploitation and exploration—engaged in by team members contribute to building team creativity, and (2) how do environmental factors—organizational learning culture, knowledge sharing, and expertise heterogeneity—affect team creativity. The results obtained using scenario-based simulations show that a direct relationship exists between team creativity and exploitation, exploration, organizational learning culture, knowledge sharing, and expertise heterogeneity. Our approach differs from the one in [11], which establish dependencies between team creativity and some specific factors. Thus, we use Bayes classifiers to build the most creative and innovative groups based on particular values of some individual characteristics (related to creativity) of the group members.

3 A Bayes Classifier-based Model and Method for Building Optimally Creative and Innovative Groups

3.1 Bayesian Network Classifiers

Classification is very important in many domains, for example in applications for object recognition (forms, human faces, characters, etc.), detecting spam e-mails or intruders in computer networks, and so on. The concept of a classifier is seen often as a correspondence between a data set (values of attributes or features of an object) and a class (category) to which the object belongs [12,13].

Formally, a classifier is defined as follows. Given a set of attributes $\{A_1, A_2, \ldots, A_n\}$with finite domains and C the class variable, also with finite domain, that corresponds to possible classes, a classifier is a correspondence $f: A_1 \times A_2 \times \ldots \times A_n \rightarrow C=\{c_1, c_2, \ldots, c_m\}$. An object is described by the values of the considered attributes $(v_1, v_2, \ldots, v_n)$ and a classifier $f(v_1, v_2, \ldots, v_n)=c_i$ shows the class to which the considered object belongs.

A Bayesian (Belief) Network (BBN) is a graphical model based on probabilistic directed acyclic graphs that can be used for representing uncertain knowledge and reasoning techniques. Each node of the graph represents a random variable, while the arcs define a probabilistic dependency between variables. These dependencies are quantified by the conditional probabilities between variables.

One of the high performance classifiers with regard to prediction of the class to which a particular object pertain are naïve Bayes classifiers [12, 14]. In case of naïve Bayes classifiers, each attribute has only one parent, namely the class variable. Naïve Bayes classifiers use Naïve Bayes Structures and Bayes' rule to predict the most probable class to which an object pertain based on the training data set. Bayes' rule is used to calculate the probability that some object pertain to a class as it is shown further on. Given an object "o", a naïve Bayes classifier estimates the probability that the object belongs to each class c_k $P(c_k|v_1, v_2, \ldots, v_n)$ to find the maximum value, according to Bayes's rule:

$$P(C=c_k|v_1, v_2, \ldots, v_n)=P(C=c_k) * P(v_1, v_2, \ldots, v_n| C=c_k)/ P(v_1, v_2, \ldots, v_n) \quad (1)$$

Using the chain rule P(v1, v2, … ,vn| C=ck) can be written as:

$$P(v_1, v_2, \ldots, v_n| C=c_k) = P(v_1| v_2, \ldots, v_n, C=c_k) * P(v_2| v_3, \ldots, v_n, C=c_k) * \\ * \ldots * P(v_{n-1}| v_n, C=c_k) * P(v_n| C=c_k) \quad (2)$$

The naïve assumption is that the attributes are independent given the class: an attributes vi is independent of attribute vj for i<>j given de class ck. Thus, the following relations are true:

$$P(v_i| v_{i+1}, \ldots, v_n, c_k)= P(v_i|c_k), \; P(v_1, v_2, \ldots, v_n| c_k)=\Pi_i P(v_i|c_k) \quad (3)$$

Given the input, $P(v_1, v_2, \ldots, v_n)$ is constant. So, the classification rule for a new object is described by $(v_1, v_2, \ldots, v_n)$ is as follows:

C^{new}<-argmax $_{ck}$ P(C=c_k) * Π_i P(v_i|c_k),where C^{new} is the estimated class. (4)

A detailed description of the classification techniques based on Bayesian Networks can be found in [12, 13, 15, 16].

3.2 General Model for Building "The Best" Collaborative Groups

In this sub-section, we introduce our model for building "the best" groups of people from a cohort of individuals. The best could mean the most creative, the most innovative, the most effective, the most proficient, etc. The main idea is to take into account a group characteristic that is the most relevant for the proposed "the best" goal and to maximize it by grouping and re-grouping people based on the values of some particular individual characteristics. Our model includes three stages and is shown in Fig. 1.

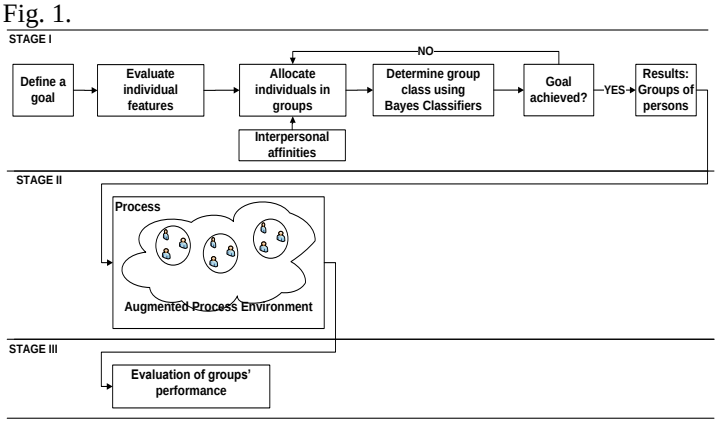

Fig. 1. Model for building "the best" collaborative groups.

Stage I

The algorithm that distributes the individuals in increasingly better groups with regard to the relevant characteristic considered includes the following steps:

- Allocate each learner to a group (ad-hoc or clique based);
- Determine the class of each of the resulted groups using Bayes classifiers;
- *If* the obtained group classification satisfies the proposed goal (for example, given 5 groups, 2 of them are in class High, 2 are in class Medium, and 1 is in class Low) then the distribution stops;
- *Else* another trial is undertaken by various combinations such as: first between the group members of the class L, second between the members of the classes L and M, and, in the end, a total re-distribution of all people.

Stage II

Within the second phase, a particular process takes place (i.e. working, learning, competing, etc.). The environment in which the process takes place can provide for achieving "the best" goal. For instance, in case of aiming at building increasingly

creative learning groups, the creative contextual learning environment can be augmented with creativity triggering activities such as: promoting the importance of creativity - learners have to be aware of creativity's role in education and in everyday life, including motivation tasks and advising tasks, using different instructional strategies (mainly the ones focused on problem-based learning and project-based learning), providing for development of social and collaborative skills, developing various teaching and learning scenarios using critical thinking models, allowing questions sessions, not over-structuring the lessons or lectures, keeping a balance between the learner control and the machine control regarding the management of the learning process, designing multicultural and multidisciplinary tasks, and including information aggregation tasks [17].

Stage III

In this stage, the performance of each group, as a whole, is assessed. For example, to evaluate group creativity, in general, the four scales defined in Torrance Tests of Creative Thinking (TTCT) may be used to measure the following aspects [18]: *fluency*: the total quantity of interpretable, meaningful, and relevant ideas produced in response to the stimulus; *flexibility*: the number of different sorts of pertinent responses; *originality*: the statistical rarity of the responses; *elaboration*: the quantity of detail in the responses. Similarly, in learning processes, an instructor evaluates the learning outcomes (ideas, products, solutions, etc.), along with each group's approach. The obtained group creativity may be low, medium or high. If "the best" goal is achieved, then the objective of constructing the most creative teams has been fulfilled, otherwise, during the next instructional session, the groups will be re-organized.

The data gathered in all the three stages are stored for further use as training data.

3.3 Model Instantiation for Building the Most Creative Learning Groups

To instantiate the model presented in the previous section, we evaluate first a set of learners' characteristics that are known to have an impact of group creativity, such as *individual level of creativity, personal motivation, domain expertise*, or any other factors that influence creativity and that we can measure. The Naïve Bayesian structure for these three attributes is shown in Fig. 2.

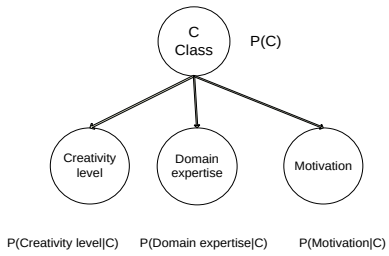

Fig. 2. The Naïve Bayes structure for groups classification.

The level of creativity may be established using various tests such as tests of divergent thinking, creative personality, etc. A well-known and easy to use test is Gough's Creative Personality Scale [19], which output range is between -12 and 18. Domain expertise is determined by assessing specific knowledge and skills (it ranges between 1 and 10). The intrinsic motivation level is evaluated using a specific questionnaire, which may result in 0 (low motivation), 1 (medium motivation), and 2 (high motivation). The Bayesian Classifier makes predictions starting with a training data set, which is obtained by performing several experiments during educational processes over long periods of time. Learners are grouped and re-grouped repeatedly until the objective with regard to obtaining the most creative groups is achieved. From the training data set, the classifier determines the conditional probability for each attribute of each individual pertaining to a certain class. Then, by applying the Bayes' rule, the probability of falling within one class for a given set of attributes is computed. The class with the highest posterior probability is the predicted class.

4 Case Study – a Real-World Scenario

We used the method and the algorithm presented in the previous section to perform a real-world grouping of students in "the best" creative and innovative teams possible with 20 of our third year Computer Science students enrolled in the Software Engineering course. The data presented here have been collected and processed during a period of five months, however, we have grouped students this way for several years now (of course, only the ones willing to participate in this educational scenario, while the others have grouped themselves either on cliques or ad-hoc).

Our "the best" goal was to obtain at least three groups with creativity class medium or higher. The learning achievements are assessed by the grade obtained in the Software Engineering course, which is based on several criteria related to domain expertise, to soft skills achieved, to the creativeness and innovativeness of the solutions, etc. The final grade measures both how well they have achieved the course requirements with respect to the domain knowledge and how well they work together in small developers' teams that aim to complete a common software development project and to properly present their work and the final product.

First, we performed a Gough-based evaluation of creativity of our students and we obtained a creativity score distribution that is presented in Fig. 3. The creativity score mean is 2.55.

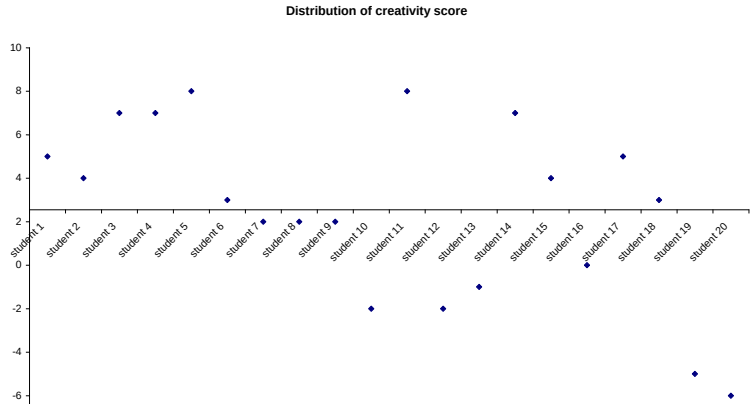

Fig. 3. Distribution of the creativity score of sophomores.

The values obtained for the two other attributes considered in the classification are shown in Table 1. The domain expertise is the grade obtained at the Data Structures and Algorithms class, while they were freshmen. We have chosen this grade because the programming part of the Software Engineering project consists of developing computer applications with fundamental data structures and algorithms in Java. The motivation attribute has been determined using an adapted questionnaire based on MSLQ, which is a multi-item self-report Likert-scaled instrument designed to assess motivation and use of learning strategies by college students [20].

Table 1. Sophomores' attributes used in classification

Learner ID	Gough score	Domain Expertise	Motiva-tion	Learner ID	Gough score	Domain Expertise	Moti-vation
Learner 1	5	8	2	Learner 11	8	10	1
Learner 2	4	8	1	Learner 12	-2	7	1
Learner 3	7	8	2	Learner 13	-1	6	2
Learner 4	7	10	2	Learner 14	7	7	1
Learner 5	8	8	1	Learner 15	4	8	1
Learner 6	3	8	2	Learner 16	0	5	2
Learner 7	2	7	0	Learner 17	5	5	2
Learner 8	2	6	0	Learner 18	3	5	0
Learner 9	2	6	1	Learner 19	-5	6	0
Learner 10	-2	5	0	Learner 20	-6	6	0

During the experiment, the students have grouped themselves in small cliques based on their inter-personal affinities (they were buddies). Four uneven cliques resulted this way (learners' IDs are presented): (1, 2, 3, 4, 5, 6), (7, 8, 9, 10), (11, 12, 13, 14, 15, 16, 17), and (18, 19, 20). After a learning session, the creativity of each group has been assessed with the following results: no team was in the high creativity class, two teams had medium creativity (coded with value 2), and two teams had low creativity

(value 1): groups 1 and 3 had medium creativity and groups 2 and 4 low creativity. As our proposed goal was to have *at least three teams* with creativity class medium or higher, we used the method presented here to re-group the students. For each student, a creativity class has been predicted using a Naïve Bayes classifier (nb1) trained with a predefined data set, as it can be seen in Fig. 4 (a Matlab cropped screenshot). For each student the values obtained, respectively are as follows: 3, 2, 3, 3, 2, 2, 2, 1, 1, 1, 3, 1, 1, 2, 2, 1, 1, 1, 1, 1.

```
>> nb1=NaiveBayes.fit(training_nb1, class_nb1);
>> cpre1 = predict(nb1,date_nb1);
>> disp(cpre1);
```

Fig. 4. Naïve Bayes classifier for individual creativity

Working our way to reaching the established goal, we re-grouped the students so that at least three of five students were buddies (this having a positive effect on team creativity in our experience and in the literature [21]). If a minimum of three of the five students are buddies, a clique is formed and, consequently, the clique attribute in Fig. 6 is set to 1 (column 6 in the left screenshot). As our objective *to obtain at least three groups with creativity class medium or higher* has not been achieved we continued further the grouping process. To determine each group's creativity a second classifier nb2 has been used. The creativity classes for the obtained groups are presented in Fig. 5 (right screenshot) - group1=(1, 2, 4, 9, 18), group2=(11, 12, 14, 15, 19), group3=(3, 5, 7, 8, 13), and group4=(6, 10, 16, 17, 20). In the left screenshot, on the first line, the data of team 1 may be read as follows: in the first column, the influence class of the student 1 to the team creativity, resulted from the Bayes-based classification on the Gough score, the domain expertise, and the motivation; in the following columns, 2 to 5, of line 1 we find the influence classes for the students 2, 4, 9, and 18. The next lines include the data of teams 2, 3, and 4, respectively.

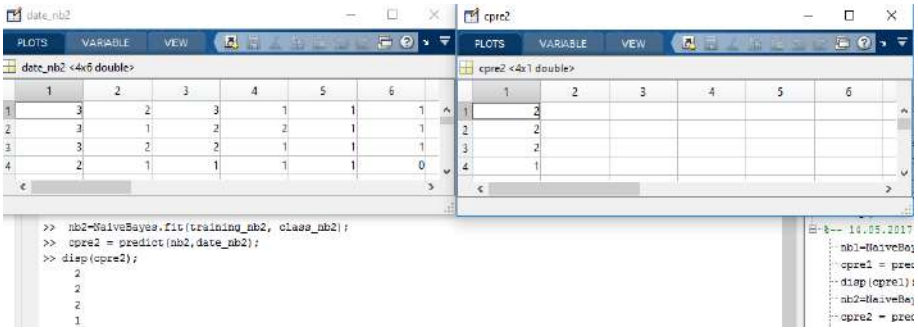

Fig. 3. Left screenshot: the influence classes for each student;
Right screenshot: the creativity classes for the groups 1 to 4

The algorithm stops because "the best" goal has been achieved (i.e. at least 3 teams have creativity medium or higher) (Fig. 6).

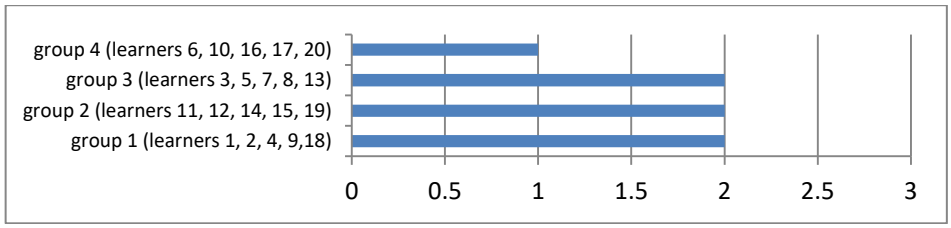

Fig. 6. Creativity classes for the groups

At the end of the first stage of our method, the groups obtained fulfill "the best" goal established. During the second stage, the learning process takes place within the augmented environment, while in the third stage the final assessment of each group's creativity takes places. In our case, the performance of the groups is measured by the grade obtained in the Software Engineering course, which is granted based on several criteria that measure the performance of each group as a whole, taking also in consideration each individual contribution. These criteria assess the developed software, the related documentation, the difficulty of the problem, the creative and innovative solutions used during development and for the presentation of the final product, the complexity of the algorithms, the cost-effectiveness of the solution and so on. As it can be seen below in Table 2, the performance of the majority of students is higher after this collaborative learning experience (first average grade is in Data Structures and Algorithms course, while the second is in the Software Engineering one). And this is not just an isolated situation, as we have already performed this grouping, in similar circumstances, for 5 years now, and the results are consistent and show increased learning with respect to both domain expertise and soft skills achieved.

Table 2. Evaluation of learning achievements before and after the grouping

Group	Average Grade DSA	Average Grade SE
Group 1	7.40	8.20
Group 2	7.60	9.40
Group 3	7.00	8.40
Group 4	5.80	8.00

5 Discussion

In order to make our technique easier to use, we explain here further some of the particularities of the method. For the time being, we use as classifying technique the naïve Bayes classifier, but some other similar classifiers may be used (for instance neural network based classifiers, decision trees, and support vector machines). The main shortcoming is presuming that the characteristics taken into account are independent of each other, which allowed us to use Naïve Bayesian Network based classifiers. This assumption may simplify things, but it can be false sometimes. Even

in our case, there can be dependencies between the individual creativity level and the domain expertise or between the motivation and the domain expertise. In the case study presented here, we have modeled the interactions inside groups using cliques. In our testing, a group was been considered to be a clique if 3 out of the 5 members of a group were buddies.

As for the first results, it can be noticed that our students in the experiment fall in the following classes with respect to creativity: 4 are in the high class (value 3 in Fig. 5), 6 are in the medium class (value 2), and 10 in the low class (value 1). This observation lead us to the conclusion that in some particular situations, when the individual creativity scores or the other two attributes are low, some bold goals cannot be achieved whatsoever. This understanding is something to consider for any kind of collaborative activity.

To experiment further, we plan to include in the parameter set some sort of evaluation of the potential for creativity obtained by external observation of both instructional and extra-curricular activities.

6 Conclusions and Future Work

Nowadays, amazing technological progress and tremendous changes in the global economy have both contributed massively to a paradigm shift with regard to collaboration among people. Increasing the efficiency and effectiveness of groups of individuals performing together specific activities to achieve common goals in given contexts is of crucial importance. Nevertheless, building high performance groups is quite challenging regardless the domain they activate it. We presented here our work on clustering individuals in groups so that a global objective with respect to the quality of groups is achieved. The grouping is based on both individual characteristics and inter-personal interactions. The proposed model and method are general and can be used for several collaborative activities such as working, learning, competing, etc. To test them, we instantiated and used them to construct the most creative and innovative collaborative learning groups. After using this method in several learning situations, we have learned that trying to semi-automatically group individuals in high performance teams with regard to some particular objectives is a laborious task that involves knowledge and instruments in various fields such as education sciences, social and personality psychology, computer science, and machine learning.

Further work ideas, besides the ones presented in the previous section, include using several creativity evaluation scales, adding contextual and organizational factors, testing the method in other activities and in other fields, improving the algorithm, performing experiments with control groups, and, eventually, offering the resulted technique as an online open service.

References

1. Johnson, D. W., Johnson, R.T., Holubec, E. J.: The New Circles of Learning: Cooperation in the Classroom and School. Assoc. for Supervision and Curriculum Development (1994).

2. Johnson, D.W., Johnson, R. T.: Making Cooperative Learning Work, Theory into Practice, 38 (2), Building Community through Cooperative Learning, 67-73 (1999).
3. Felder, R. M., Brent, R.: Cooperative Learning. In: Mabrouk, P. A. (ed.) Active Learning: Models from the Analytical Sciences, ACS Symposium Series 970, 34–53, Washington, DC: American Chemical Society (2007).
4. Felder, R.M., Brent, R.: Effective Strategies for Cooperative Learning. J. Cooperation & Collaboration in College Teaching 10(2), 69–75 (2001).
5. Moise, G., Vladoiu, M., Constantinescu, Z.: GC-MAS - a Multiagent System for Building Creative Groups used in Computer Supported Collaborative Learning. In: 8th International KES Conference on Agents and Multi-agent Systems – Technologies and Applications, Advances in Intelligent Systems and Computing, 296, 313-323 (2014).
6. Zhou, C., Luo, L.: Group Creativity in Learning Context: Understanding in a Social-Cultural Framework and Methodology. Creative Education 3(4), 392-399 (2012).
7. Reiter-Palmon, R., Wigert, B., and de Vreede, T.: Team Creativity and Innovation: The Effect of Group Composition, Social Processes, and Cognition. In Handbook of Organizational Creativity, Elsevier Science & Technology (2011).
8. Paulus, P. B., & Dzindolet, M. T.: Social influence, creativity and innovation. Social Influence (3), 228–247 (2008).
9. Gerhard Fischer, G., Giaccardi, E., Eden, H., Sugimoto, M. and Ye, Y.: Beyond Binary Choices: Integrating Individual and Social Creativity. International Journal of Human-Computer Studies - Computer support for creativity 63(4-5), 482-512 (2005).
10. Kenny, A.: 'Collaborative creativity' within a jazz ensemble as a musical and social practice, Thinking Skills and Creativity 13, 1-8 (2014).
11. Choi, D. Y., Lee, K. C. and Seo, Y. W.: Scenario-Based Management of Team Creativity in Sensitivity Contexts: An Approach with a General Bayesian Network. In: Lee, K. C. (Ed.), Digital Creativity Individuals, Groups, and Organizations (2013).
12. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning (29), 131–163 (1997).
13. Jensen, F.V., Nielsen, T. D.: Bayesian Networks and Decision Graphs. Springer-Verlag New YorkMedia (2007).
14. Duda, R. O., Hart, P. E.: Pattern Classification and Scene Analysis. Wiley (1973).
15. Pearl, J., Russel, S., Bayesian networks. In: M. Arbib (Ed.), Handbook of Brain Theory and Neural, Technical Report Networks, MIT Press (2001).
16. Störr, H. P.: A compact fuzzy extension of the Naive Bayesian classification algorithm. In: Phuong, N. H., Nguyen H. T., Ho, N. C., and Santiprabhob, P. (eds.), Proceedings InTech/VJFuzzy'2002, pp. 172-177 (2002).
17. Moise, G.: Fuzzy Enhancement of Creativity in Collaborative Online Learning. In: Chiu D. K.W. et al. (eds.), LNCS 7697, pp. 290-299 (2014).
18. Torrance, E. P.: Torrance tests of creative thinking. Scholastic, Testing Service, Bensenville, IL (1966).
19. Gough, H. G.: A Creative Personality Scale for the Adjective Check List. Journal of Personality and Social Psychology (37), 1398-1405 (1979).
20. Pintrich, P. R., Smith, D. A. F., Garcia, T., McKeachie, W. J: Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). In Educational and Psychological Measurement, 53(3), 801-813 (1993).
21. Mueller, J., Cronin, M. A.: How Relational Processes Support Team Creativity. In: Mannix, E. A., Goncalo, J. A., Neale M. A. (eds.): Creativity in Groups, Research on Managing Groups and Teams Series. 12, Emerald Group Publishing Ltd, 291-310 (2009).

Data: 5 mai 2025