A Deep Learning Approach to Estimate Multi-Level Mental Stress from EEG using Serious Games

Joaquin J. Gonzalez-Vazquez, Lluis Bernat, Jose L. Ramon, Vicente Morell, and Andres Ubeda, *Member, IEEE*

***Abstract*—Stress is revealed by the inability of individuals to cope with their environment, which is frequently evidenced by a failure to achieve their full potential in tasks or goals. This study aims to assess the feasibility of estimating the level of stress that the user is perceiving related to a specific task through an electroencephalograpic (EEG) system. This system is integrated with a Serious Game consisting of a multi-level stress driving tool, and Deep Learning (DL) neural networks are used for classification. The game involves controlling a vehicle to dodge obstacles, with the number of obstacles increasing based on complexity. Assuming that there is a direct correlation between the difficulty level of the game and the stress level of the user, a recurrent neural network (RNN) with a structure based on gated recurrent units (GRU) was used to classify the different levels of stress. The results show that the RNN model is able to predict stress levels above current state-of-the-art with up to 94% accuracy in some cases, suggesting that the use of EEG systems in combination with Serious Games and DL represents a promising technique in the prediction and classification of mental stress levels.**

***Index Terms*—EEG, Serious Games, Deep Learning, Mental stress**

I. Introduction

STRESS is one of the most common conditions in modern society, and its impact on the health and well-being of individuals can be profound. Each person perceives and manages stress to a greater or lesser extent, where poor management can lead to chronic stress, which has a demonstrable relationship with physical and mental health problems [1], [2]. The brain is intimately connected to stress. When a person experiences a stressful event, triggers a response that involves the release of hormones and the bodily adaptation in order to overcome that situation [1], [3], [4]. In those cases, they may confront these situations with varying levels of confidence, tension, fear or tranquility [5], [6], [7].

In recent decades, the study of stress has gained significant importance in several fields such as medicine, psychology, education or professional activity, among others. Thanks to technological and scientific advancements, a variety of techniques and tools have been developed to measure and analyze stress. Furthermore, artificial intelligence (AI) and machine learning (ML) have enabled the development of predictive and classification models for stress, opening up new possibilities in research and treatment of this condition [8]. A way of evaluating stress is recording and processing EEG signals from the surface of the scalp using non-invasive electrodes [9], [10].

J.J. Gonzalez, L. Bernat, J.L. Ramon, V. Morell and A. Ubeda are with the Human Robotics Group, University of Alicante, Spain. Corresponding author: andres.ubeda@ua.es

A number of studies have emerged focusing on the extraction of relevant information from human EEG signals, with their value lying in their application to the diagnosis and treatment of a wide range of neurological and psychiatric disorders [10], [11], [12].

When addressing the evaluation of stress, a few recent examples of ML approaches can be cited. For instance, Kalas et al. conducted a study on stress detection using EEG signals and applied k-means clustering to classify stress-relax binary states. The study compared this classification with an objective assessment based on physiological variables [13]. Another work by Perez-Valero et al. quantitatively evaluated stress through virtual reality (VR), measuring brain activity of 25 participants. Their study employed individualized machine learning models based on regression algorithms and achieved a high correlation with ground-truth stress levels [14]. Further analysis with Deep Learning (DL) models for stress detection has also been developed due to improved computational capacity of this algorithms. The study of Pandey et al. focused on emotion detection, including stress, using DL and categorizing emotions into valence and arousal. They applied a deep neural network (DNN) model with approximately 60% accuracy on test data. Comparative analysis with classical machine learning classifiers demonstrated similar or superior results depending on the specific application [8], [15]. Finally, a study conducted by Xu et al. addressed the use of DL combined with Transformers. This work shows promising results of up to approximately 92% accuracy [16].

All previous approaches share the common goal of stress detection, each employing different techniques such as ML and DL, but stress induction lacks of interactivity. Although there are some studies that employ interactive approaches to measure mental stress in many cases they only rely on passive audiovisual experiences as a method of stress monitoring [17]. The present study proposes an innovative differentiating approach based on Serious Games. This concept is linked to games developed for purposes other than entertainment, aiming to provide value and utility in their focused field. Our proposal employs this kind of games as an interactive method for the quantitative measurement of mental stress, seeking greater user involvement and attention during the monitoring process with the objective of corroborating if this interactive approach is an effective method for the prediction and classification of mental stress from EEG signals. For the identification of the stress level, we propose a classification model consisting of a RNN based on a GRU architecture, a variant of the Long Short-Term Memory (LSTM) network, due to the time-series nature of the recorded signals.

II. Materials and Methods

The proposed system consists of a Serious Game about a racing game focused on avoiding obstacles with a vehicle controlled by the side arrow keys on the computer keyboard (see Figure 1). The game is divided into 4 levels of difficulty, each differentiated by the increasing number of obstacles to dodge. As described in the introduction, the use of the game as the final application is based on the interactivity and attention it generates in users during gameplay. An 8-channel EEG device is used as the recording method to capture the brain signals of a group of 19 users, from which a dataset of signals is obtained based on the played level, repetition, channel, and user. The dataset undergoes a preprocessing stage to optimally input the data into the classification model. The goal is to classify mental stress into 4 levels of intensity correlated with the 4 levels of difficulty in the Serious Game. The DL model consists of RNNs with a structure of 7 layers based on GRUs. The model is trained with the preprocessed dataset to tackle the multi-class classification problem of mental stress into 4 intensity levels.

A. Serious Game

The development of a game is a complex and creative process that involves multiple stages, from the conception of the idea to its production. The development process for the proposed game takes place in the Unity game engine, which brings together the necessary tools to import materials, textures, and effects that shape the game and allow for the construction of the application to be integrated into the EEG system.

The initial idea of the game addresses how to make the player face various situations that generate different levels of stress. To achieve this, based on the concept of stress-generating Information and Communication Technologies (ICTs) factors, a racing game with multiple difficulty levels is proposed. In each level, the number of obstacles to dodge will increase. This idea is based on the ICT aspect of information overload, where the increase in visual elements (obstacles) in each level creates an overload that induces tension and stress. Additionally, technical difficulty is induced by creating four levels, thereby increasing the game's challenge. On the other hand, social isolation is induced by requiring the player to perform the test alone without distractions, with the car engine and game sound isolating external sounds [18], [19].

The game presents randomly generated roads and obstacles, the car moves by scrolling vertically through these random roads during the level. The difficulty of each level varies depending on the increase or decrease in the number of obstacles (trucks) placed on each road. Players can change the horizontal position of the car while it moves at a constant velocity through the screen. A countdown at the beginning of each level has been added together with sounds activated when the car crashes with the obstacles which are represented by long yellow trucks. When a player crashes with an obstacle it disappears and an explosion sound is played to increase the sense of failure. Additionally, Play Now and Play Again buttons have been added for better management.

Additionally, a survey has been conducted to ensure that the game levels generate a stress level comparable to the difficulty of each level (Figure 2). Each participant was asked to decide which amount of stress (from 1 to 10) was feeling while playing each of the game levels. The survey shows that each of the 4 levels provide a clear difference in stress perception.

B. EEG System and Communication

The EEG system used is the Unicorn Hybrid Black acquisition device (G.tec Medical Engineering GmbH, Austria) connected to a laptop computer with the final application of the Serious Game programmed in Unity and the recording API programmed to integrate the game with the acquisition device. The Unicorn Hybrid Black is an EEG device that allows the measurement of brain electrical activity in different scalp areas through 8 surface electrodes located following the 10/20 International System (Fz, C3, Cz, C4, Pz, Po7, Oz, P08). The EEG signals are sampled at 24 bits and 250 Hz per channel, and referenced to two electrodes placed on the mastoids [20]. This device has been chosen to evaluate if mental stress detection is feasible even with a low-cost device and outside the laboratory environment.

For the integration of the Unicorn recording device and the Serious Game, an API is programmed using UDP sockets to enable communication between the two parts. The recording is programmed to last exactly 90 seconds.

C. Experimental Protocol

Brain signals have been measured from a group of 19 users (15 men and 4 women with an average age of 31 ± 14). 12 of the participants are right-handed and 7 are left-handed. All participants have signed the correspondent informed consent following the Ethics Committee Protocol (REF: UA-2023-02-08_2) according to the Declaration of Helsinki, which establishes the postures, breaks, and timing to be followed. During the test, each user is monitored by level and repetition, with 5 repetitions for each of the 4 difficulty levels of the Serious Game, resulting in 20 EEG measurements of 90 seconds for each of the 19 participants, generating a 30-minute EEG recording per user.

D. EEG Signal Preprocessing

Once the brain signals of each user are captured, a preprocessing window is established with the aim of cleaning, removing noise, and transforming the data to arrange it optimally for subsequent classification. To achieve this, a low-pass filter is applied to eliminate higher frequencies caused by motion artifacts or high-frequency electromagnetic interference, allowing only frequency components up to 60 Hz to pass through. A high-pass filter is also used to remove lower frequencies present in the signal caused by slow-motion artifacts such as head position changes, slow eye movements or finger movements during the game interaction. The applied filter removes frequencies ranging from 0 to 4 Hz to eliminate

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3395548

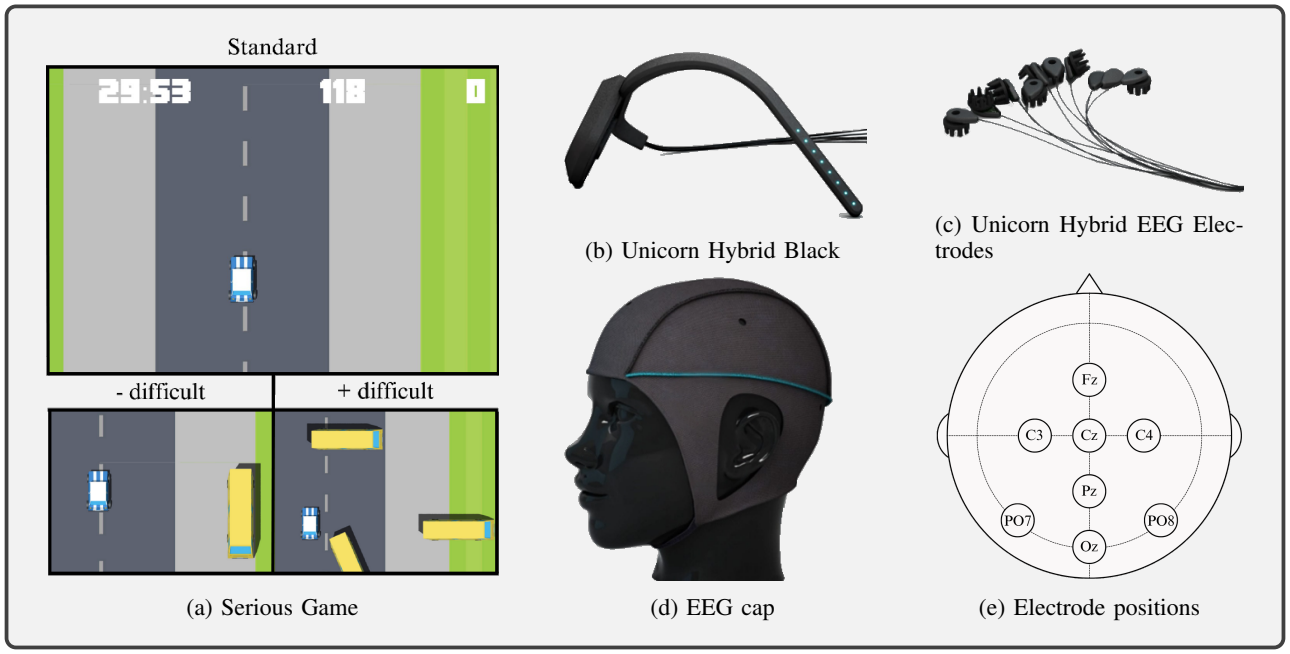

Fig. 1: Experimental setup composed of the Serious game (a), and the EEG system composed of the LED array (b), electrodes (c), cap (d) and standardized channel locations (e).

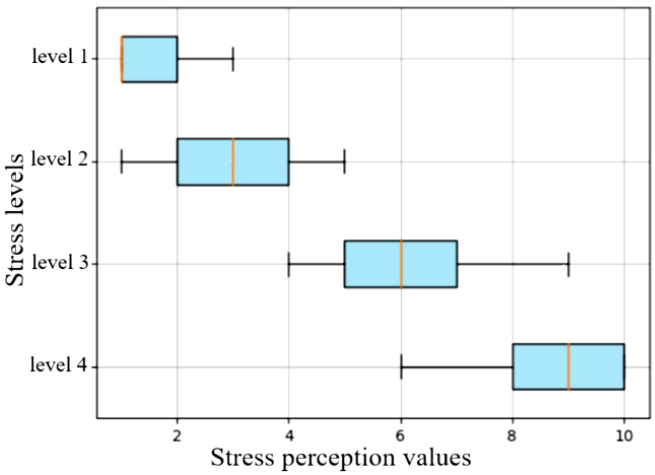

Fig. 2: User's subjective stress perception at each difficulty level. 1-10 Likert Scale (n=19).

all types of motor and visual artifacts present during the experiment. A Notch filter is implemented to eliminate the 50 Hz component caused by the power grid interference. Additionally, an adaptive averaging filter is applied to remove residual noise or interference by adapting to the input signal while preserving the relevant signal features. Other methods, such as Independent Component Analysis (ICA), have been discarded due to the limited number of channels and uncertain reliability.

In order to generate a larger amount of data to train the DL model, the EEG signal from the 8 channels is divided into 2-second segments with a 20% overlap. This has been done to generate a large amount of data for the neural network, which also allows for the generation of a greater number of patterns for the network to learn from and enhances computational efficiency. The generated dataset is stored in a matrix saved in a MATLAB (.mat) file, organized in folders by user. Each file is labeled according to the attempt and difficulty level of the game.

The remaining transformations and preprocessing steps take place in Kaggle (online community of ML). The Kaggle's environment used to preprocess and also train the models provides a Tesla P100 GPU with 16 GB of RAM and weekly access of 30 hours. The data is labeled with a value ranging from 0 to 3, corresponding to the difficulty level of the respective level. The specific labeling is as follows: "0: Low stress, 1: Moderate stress, 2: Intermediate stress, 3: High stress".

Prior to preparing the dataset for the model, it is necessary to ensure the removal of as many artifacts as possible, which involves performing a second preprocessing step for this purpose. Values exceeding the absolute threshold of 75 mV are considered outliers, as they exceed the resting potential of neurons [14]. These outliers can be caused by blinks, user movements, or simply uncommon recordings of brain activity. To handle them, they are mitigated to 0 mV without being removed from the dataset.

It is worth noting that DL models perform better with data normalized between the values of 0 and 1. This is because normalized data has lower variance and is less sensitive to any remaining outliers. Additionally, normalized data provides numerical stability and allows the model to compare and extract features more easily since all the data is on the same scale. Therefore, the data has been normalized and rounded to one decimal place to avoid overloading subsequent calculations due to the number of decimal places.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3395548

Once the data has been preprocessed, the next step is to transform it to make it suitable as an input for the model. Considering that a minimum of 19 stress classification models will be implemented (one model per user), the computational cost will be high. To leverage the capabilities provided by Kaggle, the data is transformed into one dimension making it suitable for one-dimensional models. Afterwards, the dataset is converted into a tensor and divided into training, testing, and validation sets. The predefined split percentage is as follows: 20% of the total data for the test set, 20% of the remaining 80% for the validation set and the remaining for the training set.

E. Multi Stress Level Classification

Once the preprocessed and transformed dataset is ready, patterns and features are extracted using a DL model to make predictions on stress level classification.

Considering that our problem involves classifying EEG signals using DL and that classification models for EEG signals do not generalize well due to the variability in how individuals express their emotions at the brain level and the varying requirements to induce stress in each person, we will train a classification model for each user. Additionally, medium-size models (4 participants) and models with all participants will be trained to establish how well this modelling can generalize.

Regarding the implementation of the model, it is necessary to establish a consensus to ensure that each generated model is identical in terms of architecture and hyperparameters, ensuring consistent classification for each user. To achieve this, it is important to consider the type of data used as input. In this case, the signal data corresponds to a time series, so a RNN model based on GRU will be implemented (Figure 3). GRU units are a variant of LSTM networks, with the difference being that each GRU unit has two gates: reset gate and update gate, instead of the three gates (input, forget, and output) present in LSTM units. This design choice increases computational efficiency with GRU units at the expense of reducing long-term storage capacity [21]. Also, GRU has been implemented due to the favorable nature of the data as a time series and its ability for short-term memory, allowing it to retain information about important patterns to learn and discern stress levels. In contrast, simple DNN are unable to learn, according to previous analysis and Transformers require a large amount of data (much greater than the current dataset) for the model to start learning patterns properly.

To build the model, an input layer is implemented with a size corresponding to the training data that will be fed into the model. This is followed by 4 GRU layers, each consisting of 128 units, except for the last layer which has 256 units. Subsequently, a flattening layer is implemented to reduce the dimensionality of the data. Finally, the output layer is added, consisting of a dense layer with 4 units representing the 4 stress levels. The softmax activation function is applied to interpret the output as a probability distribution across the 4 stress levels [22].

To compile the model, an optimizer, a loss function, and a performance evaluation metric need to be defined. Firstly, the

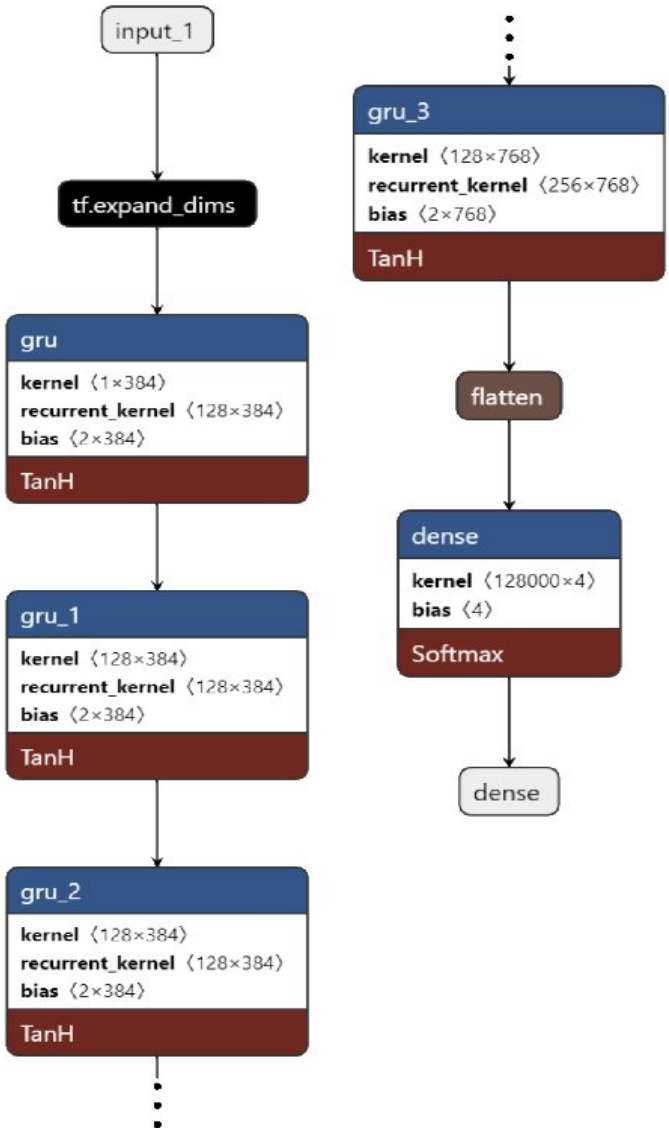

Fig. 3: DL model architecture composed of GRU layers where the input data is routed to a dense layer with ”softmax” activation function to confront the multi-class classification problem.

”Adam” (Adaptive Moment Estimation) optimizer has been implemented. This optimization approach allows the network to adapt to different learning rates and adjust the model to the specific problem at hand [22]. As for the loss function, ”categorical cross-entropy” has been implemented, as the address problem is multi-class and this function evaluates the discrepancy between the predicted probabilities and the actual classes, aiming to minimize this divergence during the model training. Additionally, a suitable performance evaluation metric should be chosen to assess the model’s performance on the task at hand. The performance evaluation metric chosen is the accuracy, which measures the fraction of correct predictions out of the total predictions made [22], [23]. Finally, the model is trained by providing the corresponding training data, which includes the signal data and their corresponding labels. The number of epochs to iterate over, the batch size of the data, and

validation data are defined. Additionally, a learning rate decay callback is introduced to gradually reduce the learning rate at each epoch. In this case, TensorFlow learning rate scheduler has been used, which adjusts the learning rate based on a lambda function that reduces the rate by a factor every X epochs. This allows for a fine-tuning of the model's learning process and potentially improves its performance.

The process of creating, compiling, and training the model is replicated for each of the 19 users and later is generalized by implementing intermediate (M) and full-dataset (G) models, resulting in a total of 25 stress level classification models. Individual models have been trained using 60 epochs. This number is selected by averaging the optimal number of epochs for each individual model. However, for larger models "M" and "G" a higher number of epochs (90) has been also selected to show if there were differences in accuracy due to the increase of this parameter.

Each model has been trained with the same hyperparameters, except for the aforementioned number of epochs and the increase in batch size in the general models, which require a longer training time to learn, specifically, being 32 for individual models, 64 for medium models, and 512 for large models. In spite of that, the architecture and size of the model remain the same. After training and validating the models, they are saved, and their performance is evaluated using unseen test data. This provides a series of results based on the metric used, accuracy. Both, the training and performance evaluation of the models, with validation and test data, are performed using a cross-validation. Each model is therefore trained and validated 5 times, to later obtain the average accuracy and deviation as the final result for the evaluation of the model.

III. Results and discussion

Results of individual (I), medium size (M) and generalized (G) models are shown in Table I, including the selected number of epochs and the mean accuracy values after classification. M1 model includes participants 1, 6, 16 and 18, while M2 includes participants 2, 4, 12 and 19. Users included in the M models are randomly selected. The G model includes all participants except for users 2, 8 and 11. These three participants were excluded from the general model because they exhibited sudden movements and unintentional gestures during the test, which contributed to the appearance of artifacts. This aspect is reflected in the individual model's accuracy of some of these users. Additionally, the letter "L" indicates that the model has been trained for a larger number of epochs.

Figure 4 shows that models I1, I4, I6, I9, I16, and I17 have achieved an accuracy greater than 90%. However, four of them, I3, I7, I10, and I11, do not reach 80% accuracy. The variability in these results is relatively low and can be attributed to factors such as noise during the test, user's sudden movements, lower concentration at certain moments during monitoring, or simply variations in the user's perception of stress levels compared to the established reference. In other words, there may be moments where the user is more or less stressed than what is defined.

Models trained with a higher number of iterations, such as M1L, M2L, or G1L, show that the increased number of epochs does not significantly improve the model's accuracy (Wilcoxon Sum Rank Test, $p > 0.05$). However, if a model were to be constructed with a very large number of participants, this would be a factor worth studying and considering.

On the other hand, the accuracy metric used in the development is one of the most suitable for this problem. Since the model's output is an integer representing the stress level, this metric can effectively show the percentage of times the model has correctly predicted a stress level. Thus, intuitively, the real performance of the model can be understood.

Theoretically, as more individuals are introduced, the accuracy of the model should decrease due to the fact that EEG data does not generalize well as it is a signal very dependent on individual brain behavior. For the intermediate models, considering that one of them was trained with individuals who had an accuracy higher than 90%, the result was lower, although the decrease was minimal compared to what was expected. With a higher number of epochs, it achieved a very similar accuracy of approximately 89%.

For the model that includes all 16 users, the reduction in accuracy was significant, reaching around 73%. However, this result is still quite promising because it was possible to generalize a model that includes a large number of different users, each with distinct psychological traits. Despite being lower than the accuracy of the individual models, which was expected, the percentage of accuracy is much higher than initially estimated. Compared to other studies, there was a possibility that with such a number of users, the model would not learn or yield a lower result.

IV. Comparative with previous studies

After presenting and discussing the results, a comparison between our study and previous works addressing stress classification from EEG signals has been done (see Table II). This comparison has been carried out only with those studies that use ML or DL as an approach for classifying mental stress level. It is worth mentioning that the disparity of protocols and approaches make more difficult this comparison but studies have been selected at least considering multi-class stress level.

Following Table II, in the study presented by Hou et al. [24] a Stroop colour-word test is used to induce different levels of stress, and two classifiers, support vector machine (SVM) and K-nearest neighbors (k-NN), are applied, comparing the accuracy results of both methods. In this study, a classification of different stress levels is made, which are divided into 2, 3 or 4 levels; this allows us to compare our model with those described in this study, seeing how our model always classifies 4 levels, in addition to the fact that the accuracy results are considerably higher in comparison for those cases.

Kalas et al. [13] conducted research on stress detection using a multi-class method such as k-means clustering to divide the dataset into 2 classes by applying a threshold. In contrast, our study segments the data into 4 classes and extracts features from 19 subjects. The evaluation metric accuracy was not used, but the classification threshold was shown as 0.3989 stress index. On the other hand, Perez-Valero et al. [14] conducted a study employing a virtual relaxation experience,

Models	Epochs	Mean Accuracy	Std
I1	60	0.912	0.012
I2	60	0.854	0.014
I3	60	0.783	0.018
I4	60	0.944	0.008
I5	60	0.866	0.012
I6	60	0.914	0.008
I7	60	0.868	0.004
I8	60	0.699	0.018
I9	60	0.897	0.008
I10	60	0.795	0.009
I11	60	0.716	0.012
I12	60	0.828	0.007
I13	60	0.874	0.005
I14	60	0.887	0.010
I15	60	0.895	0.009
I16	60	0.932	0.012
I17	60	0.939	0.006
I18	60	0.790	0.015
I19	60	0.855	0.020
M1	60	0.858	0.004
M2	60	0.841	0.007
M1L	90	0.837	0.010
M2L	90	0.844	0.016
G1	60	0.722	0.005
G1L	90	0.714	0.007

TABLE I: Results of trained models. The nomenclature for each model corresponds to the letter ”I” followed by the participant number for individual models, the letter ”M” followed by the model number for medium-sized models. Additionally, the letter ”L” indicates that the model has been trained with a larger number of epochs. Finally, the letter ”G” followed by the model number represents the large-sized model that includes all users except for three (I2, I8, I11). The columns ”Mean Accuracy” and ”Std” represent the mean and the corresponding standard deviation obtained from a 5-fold cross validation.

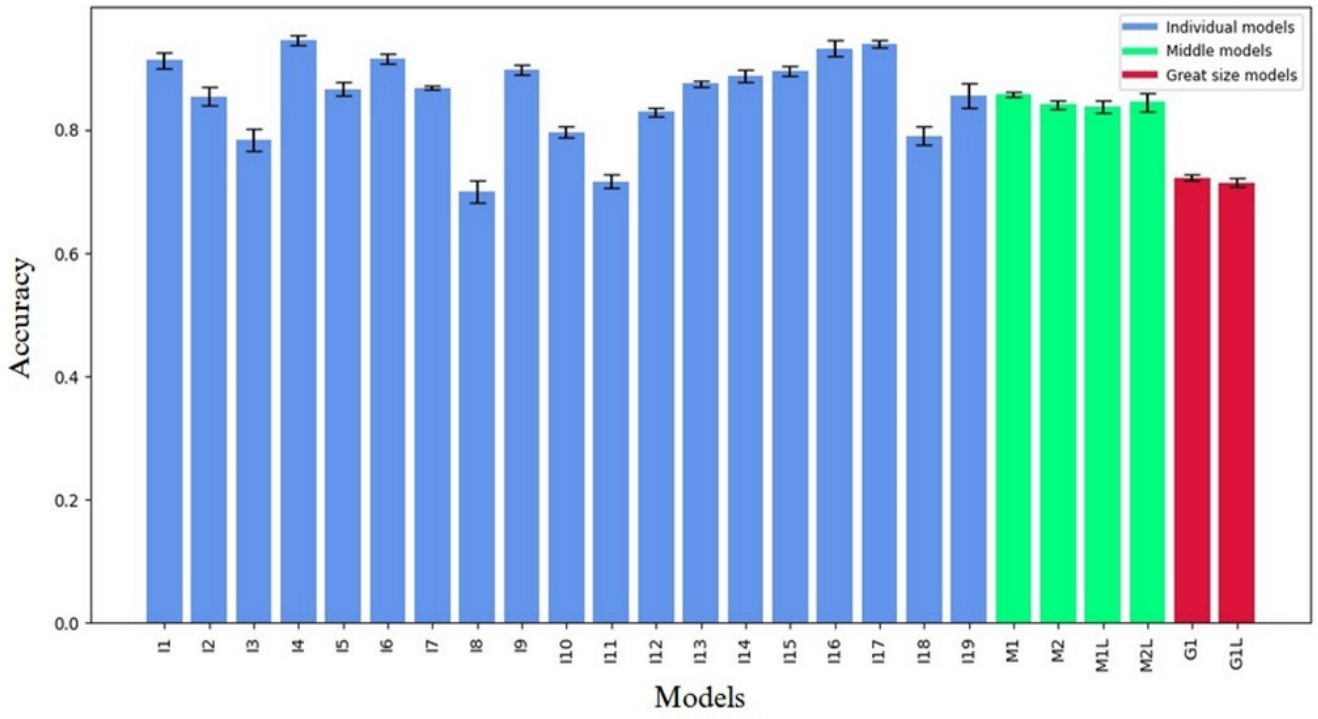

Fig. 4: Mean accuracy and standard deviation for each proposed model. I1-I19 corresponds to individual models, M1-M2L corresponds to middle-size models and G1 and G1L corresponds to generalized models.

Study		Hou et al. (2015) [24]	Perez-Valero et al. (2021) [14]	Pandey et al. (2022) [15]	**OURS**
Participants		9	23	26	19
Dataset source		Own	Own	External	Own
EEG Channels		14	64	64	8
Method to induce stress		Stroop colour-word test	VR-induced relaxation-stress states	Induction of different moods	Serious game with 4 levels of difficulty
Classes to be classified		(2, 3), 4	5	2	4
Metrics		Accuracy	Correlation	Accuracy	Accuracy
Performance	**Worst**	0.47, 0.40	0.7	0.56	0.73
	Average	$0.67\pm0.13, 0.54 \pm 0.11$	0.92	0.60 ± 0.04	0.86 ± 0.07
	Best	0.84, 0.74	1	0.62	0.95
Generalised performance (if evaluated)		–	–	–	0.73

Study		Xu et al. (2022) [16]	Roy et al. (2023) [25]	Sundaresan et al. (2021) [26]	**OURS**
Participants		23	48	13	19
Dataset source		Own	External	Own	Own
EEG Channels		64	14	16	8
Method to induce stress		Music and arithmetic operations	SIMKAP experiment	Arithmetic operations	Serious game with 4 levels of difficulty
Classes to be classified		2	2	4	4
Metrics		Accuracy	Several (we show accuracy)	Accuracy	Accuracy
Performance	**Worst**	0.62	<0.9	0.91	0.73
	Average	0.81	<0.95	0.93	0.86 ± 0.07
	Best	0.93	0.98	0.99	0.95
Generalised performance (if evaluated)		0.75	–	–	0.73

TABLE II: Summary of State-of-the-Art of stress level classification from EEG signals.

unlike the interactive "serious game" approach in this study. They extracted features from 23 participants to classify 2 states (relaxed and stressed), whereas this study includes 19 users and aims to classify 4 classes. They used classical ML and various methods for classification, unlike the DL and accuracy metric used here. Their result showed a correlation coefficient between 0.7 and 1, but this metric could have flaws as it ignores amplitude and baseline when comparing data and can lead to inaccurate predictions despite a high coefficient.

The study by Pandey et al. [15] also uses DL for stress classification, but they use a DNN instead of LSTM net. Their approach involves a 64-channel EEG device, a public dataset with 26 subjects, and the classification of emotions into two variables: valence and arousal. They achieve around 60% accuracy, lower than the current study, but find that DL is the most effective choice for their problem, compared to the tested ML method SVM. Both projects use the accuracy metric, showing the model's real effectiveness.

The study by Xu et al. [16] focuses on using Transformers with DL networks, a popular and novel approach. They use different music types for relaxation and arithmetic tasks for stress induction, monitoring 23 users with a 64-electrode EEG device, 512 Hz sampling, and 2 classes (stressed and non-stressed) with 5 sub-levels each. Their goal is similar to our study, seeking a universal method for stress measurement and classification. Results are comparable, with slightly lower accuracy at around 62% and 92%, and a generic model achieving approximately 73% accuracy.

Other studies primarily employed RNNs among other neural networks. For instance, Roy et al.'s research [25], implements LSTM, BiLSTM, GRU, and hybrid networks combining CNN with RNN. Several models were trained with an EEG signals dataset collected through the application of a subjective test on users that classifies signals into two binary states (stressed and relaxed). In comparison with their study, ours takes an interactive approach (Serious Game) as opposed to the employed test; furthermore, our study performs a multi-class classification of stress levels as opposed to their binary classification.

Another interesting study is Sundaresan et al. [26], which investigated stress in patients with Autism Spectrum Disorder (ASD). Stress was induced by subjecting patients to perform complex arithmetical calculations and an RNN was used for classification of four classes that include a stressful state and guided or unguided breathing. Similar to our study, this research presents a multi-class classification approach to stress levels. However, it lacks of an interactive approach and does not classify different stress levels.

Upon observing the table, excluding the study by Perez-Valero et al. [14] that applies a different metric based on correlation, the best-performing study is the current one, followed by the study of Xu et al. [16] with similar metrics. The study of Sundaresan et al. [26] obtains a slightly better results but it is focused on classifying a single stress state vs other conditions. The study of Pandey et al. [15] exhibits the least variability in the performance of different models, showing only 4% variation in the various tests. This indicates that despite not yielding the best results, the trained models are quite consistent, performing similarly to each other. Moreover, the studies by Hou et al. [24], by Perez-Valero et al. [14], by Sundaresan et al. [26] and the proposed in this paper are the ones with more than 2 classes to classify. This is important to consider, as an increase in the number of classes makes

it a multi-class problem, posing greater difficulty. It is worth noticing that only one study (Xu et al.[16]) performs model generalization.

Considering the relationship between the number of classes and performance, the generalization approach and the use of the accuracy metric, the current study achieved the best performance, reaching up to 0.95 for this metric. This success is attributed to the ongoing adjustments made to the architecture and hyperparameters, persistently refining them until discovering the optimal combination for achieving the highest performance levels. Additionally, the utilization of GRU played a significant role in enhancing the model's capabilities. Furthermore, the introduction of an interactive method, specifically the serious game, proved to be instrumental in inducing stress in a more nuanced and effective manner. Also, a unique dataset was created based on capturing information through EEG signals of users playing a serious game, which generates an interactive data capture environment distinguishing it from other passive datasets contexts.

V. Conclusions

In this study, an EEG system has been developed combined with a Serious Game capable of measuring stress in a relatively straightforward manner due to the intuitive control presented. This approach constitutes a distinctive system compared to conventional EEG-based systems used for stress measurement, which primarily involve the application of a visual and auditory experience where participants alternate between states of relaxation and stress. The difference lies in the user's interactivity with the game, as opposed to the passivity of traditional systems.

Concerning the multi-class prediction and classification of stress levels, several factors have contributed to the success of our proposal, such as the carefully devised measurement protocol and specific preprocessing undertaken. This has provided promising results towards the goal of implementing a comprehensive stress model. Our proposal outperforms most of the current state-of-the-art regarding general accuracy (0.95 with very high stability) and is better in terms of number of classified stress levels (4 vs the conventional binary classification of stress vs relax). This suggests that the utilization of Serious Games and DL could potentially evolve into an effective technique for mental stress classification.

Acknowledgments

The authors want to thank all the participants in the study.

References

[1] Kemeny, M. E. (2003). The psychobiology of stress. Current directions in psychological science, 12(4), 124-129.

[2] Ehrlich, M., & Mitchell, J. K. (1994). Working stress design method for reinforced soil walls. Journal of geotechnical engineering, 120(4), 625-645.

[3] Dallman, M. F. (2010). Stress-induced obesity and the emotional nervous system. Trends in Endocrinology & Metabolism, 21(3), 159-165.

[4] Wadhwa, P. D., Sandman, C. A., & Garite, T. J. (2001). The neurobiology of stress in human pregnancy: implications for prematurity and development of the fetal central nervous system. Progress in brain research, 133, 131-142.

[5] Ungerleider, L. G., & Haxby, J. V. (1994). 'What'and 'where'in the human brain. Current opinion in neurobiology, 4(2), 157-165.

[6] Park, H. J., & Friston, K. (2013). Structural and functional brain networks: from connections to cognition. Science, 342(6158), 1238411

[7] Pereira, T. D., Shaevitz, J. W., & Murthy, M. (2020). Quantifying behavior to understand the brain. Nature neuroscience, 23(12), 1537-1549.

[8] Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. Iraqi Journal for Computer Science and Mathematics, 3(1), 115-123.

[9] Wallace, B. E., Wagner, A. K., Wagner, E. P., & McDeavitt, J. T. (2001). A history and review of quantitative electroencephalography in traumatic brain injury. The Journal of head trauma rehabilitation, 16(2), 165-190.

[10] Sterman, M. B. (2000). Basic concepts and clinical findings in the treatment of seizure disorders with EEG operant conditioning. Clinical electroencephalography, 31(1), 45-55.

[11] Hyland, H. H., Goodwin, J. E., & Hall, G. E. (1939). Clinical applications of electroencephalography. Canadian Medical Association Journal, 41(3), 239.

[12] Curran, E. A., & Stokes, M. J. (2003). Learning to control brain activity: A review of the production and control of EEG components for driving brain–computer interface (BCI) systems. Brain and cognition, 51(3), 326-336.

[13] Kalas, M. S., Momin, B. F. (2016). Stress detection and reduction using EEG signals. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 471-475.

[14] Perez-Valero, E., Vaquero-Blasco, M. A., Lopez-Gordo, M. A., Morillas, C. (2021). Quantitative assessment of stress through EEG during a virtual reality stress-relax session. Frontiers in Computational Neuroscience, 15, 684423.

[15] Pandey, P., Seeja, K. R. (2022). Subject independent emotion recognition from EEG using VMD and deep learning. Journal of King Saud University-Computer and Information Sciences, 34(5), 1730-1738.

[16] Xu, X., Zhao, Y., Zhang, R., & Xu, T. (2022). Research on Stress Reduction Model Based on Transformer. KSII Transactions on Internet and Information Systems (TIIS), 16(12), 3943-3959.

[17] Badr, Y., Al-Shargie, F., Tariq, U., Babiloni, F., Al Mughairbi F., & Al-Nashash, H. (2023). Classification of Mental Stress using Dry EEG Electrodes and Machine Learning. Advances in Science and Engineering Technology International Conferences (ASET), 1-5.

[18] Day, A., Scott, N., & Kevin Kelloway, E. (2010). Information and communication technology: Implications for job stress and employee well-being. In New developments in theoretical and conceptual approaches to job stress, 317-350.

[19] Lee, A. R., Son, S. M., & Kim, K. K. (2016). Information and communication technology overload and social networking service fatigue: A stress perspective. Computers in human behavior, 55, 51-61.

[20] Breinbauer, S. (January 12, 2023). Home. Unicorn Hybrid Black. Retrieved on March 15, 2023 from https://www.unicorn-bi.com/

[21] Yang, S., Yu, X., & Zhou, Y. (2020). Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In 2020 International workshop on electronic communication and artificial intelligence (IWECAI), 98-101

[22] Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. The Journal of Machine Learning Research, 20(1), 1934-1965.

[23] Maxwell, A. E., Warner, T. A., Guillén, L. A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. Remote Sensing, 13(13), 2450.

[24] Hou, X., Liu, Y., Sourina, O., Tan, Y. R. E., Wang, L., & Mueller-Wittig, W. (2015). EEG Based Stress Monitoring. IEEE International Conference on Systems, Man, and Cybernetics, 3110-3115.

[25] Roy, B., et al. (2023). Hybrid Deep Learning Approach for Stress Detection Using Decomposed EEG Signals. Diagnostics, 13.11, 1936.

[26] Sundaresan, A., et al. (2021). Evaluating deep learning EEG-based mental stress classification in adolescents with autism for breathing entertainment BCI. Brain Informatics, 8.1, 1-12.