Stress Recognition with EEG Signals Using Explainable Neural Networks and a Genetic Algorithm for Feature Selection

Eric Pan(✉) and Jessica Sharmin Rahman

School of Computing, Australian National University, Canberra, Australia
{eric.pan,jessica.rahman}@anu.edu.au

Abstract. Stress is a natural human response to external conditions which have been studied for a long time. Since prolonged periods of stress can cause health deterioration, it is important for researchers to understand and improve its detection. This paper uses neural network techniques to classify whether an individual is stressed, based on signals from an electroencephalogram (EEG), a popular physiological sensor. We also overcome two prominent limitations of neural networks: low interpretability due to the complex nature of architectures, and hindrance to performance due to high data dimensionality. We resolve the first limitation with sensitivity analysis-based rule extraction, while the second limitation is addressed by feature selection via a genetic algorithm. Using summary statistics from the EEG, a simple Artificial Neural Network (ANN) is able to achieve 93.8% accuracy. The rules extracted are able to explain the ANN's behaviour to a good degree and thus improve interpretability. Adding feature selection with a genetic algorithm improves average accuracy achieved by the ANN to 95.4%.

Keywords: Stress detection · Artificial Neural Network · EEG · Rule extraction · Neural network explainability · Genetic algorithm

1 Introduction

1.1 Background

Stress exists for humans in all domains, whether it is work, study, or otherwise situations with external pressures. There are many other forms of stress, all of which depend on psychological factors and induce physiological responses [3]. It is imperative to have a method of measurement that can objectively quantify important symptoms or indicators of stress. This is especially the case where specialist psychologists are not available to exercise expert judgement and identify stress [4]. One tool for objective measure is the electroencephalogram (EEG). By successfully discovering patterns in EEG signals instrumental to stress recognition, our findings can provide stress researchers with more confidence on its efficacy in this domain.

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 136–143, 2021.
https://doi.org/10.1007/978-3-030-92310-5_16

Artificial Neural Networks (ANNs) are good function approximators that also excel at simple classification tasks. Despite being able to achieve high performance and good results in terms of predictions and classifications, many domain experts are skeptical to use them to make highly crucial decisions that have significant ramifications if done wrong [7]. This is because the knowledge represented in the parameters of ANNs are difficult to interpret. Unless a human can logically interpret its actions in the context of the domain, experts cannot justify it as a decision-making tool. This is prevalent in domains with high ethical stakes or where explanations must be given to key stakeholders. This disadvantage has led to the development of algorithms which extract rules and behavior patterns from neural networks, which are easy for humans to understand [6].

Another prominent issue in the world of machine learning is high-dimensional data. A large number of features gives rise to the problem of data sparsity, and it becomes difficult for models to generalize and learn useful patterns. This issue is prevalent in tasks involving EEG data since there are so many channels [9]. Feature selection therefore becomes extremely important when working with high dimensional data.

1.2 EEG Signals

The EEG is a commonly used medical imaging technique which reads electrical activity from the surface of the scalp generated by the brain. These readings are human physiological features which undergo change when a person experiences different emotions, including stress [12]. Combined with high temporal resolution (large reading frequency) [12] makes the EEG an ideal tool for stress detection. The signals used in this paper come from a 14-channel headset. Each channel detects activity from a different part of the brain. These channels are: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. Such contact-based devices are common amongst studies involving physiological reactions, such as classifying emotions [1].

1.3 Proposed Task

We aim to perform a binary classification on stress. The goals of the paper are to demonstrate the merits of using wearable devices to learn about stress, to provide confidence that neural networks can be explained intuitively, and to show how the right kind of input processing can dramatically yield better results. We leverage the predictive ability of neural networks to do this, while deducing meaningful rules that compress the neural network's behavior into a digestible, explainable series of decisions. To select only the useful features, genetic algorithms (GA) are among the methods that can be configured freely with parameters to improve efficacy [10].

The experiment will be conducted in two phases. The first phase includes building the optimal ANN architecture for the EEG dataset, manually selecting features qualitatively, and then implementing the sensitivity analysis-based rule extraction for the network. The second part will be identical to the first except

the optimal features are selected by the genetic algorithm. Their respective performances are compared. The rest of this paper is structured as follows. Section 2 describes the methods and techniques used in our study, including the dataset, architecture, rule extraction and feature selection. Section 3 presents the performance results of these techniques in identifying stress. Finally, Sect. 4 discusses future work and concludes the paper.

2 Methodology

2.1 Data Exploratory Analysis and Preprocessing

For this study, 25 undergraduate students wore EEG devices while watching a series of stressed and non-stressed films lasting around one minute. Stressed films had stressful content in the direction towards distress, fear and tension whereas the non-stressed films had content that created an illusion of meditation or soothing environments. There were three of each category of film. We note that stress can certainly be measured on a sliding scale, however in this experiment, the stimuli to induce stress has only two levels. For each of the 14 channels (described in Sect. 1.2), raw measurements were taken over time for each participant, forming time series data. The following summary statistics of the time series data were produced for each EEG channel: mean, min, max, sum, variance, standard deviation, interquartile range, skewness, root mean square, averages of first and second differences between consecutive signal amplitudes, Hjorth mobility parameter and Hurst exponent. The approximate entropy and fuzzy entropy measured randomness in signal fluctuations. Each observation contains a label of *calm* (1) or *stressed* (2) according to the category of film shown.

In the first phase of the experiment, features selection was done manually by identifying variables which are intuitively redundant. In the second phase of the experiment, we implemented a genetic algorithm to stochastically search for the optimal subset of features to be used by the ANN. We discuss this method in detail in Sect. 2.3.

2.2 Neural Network Design

The optimal model was found to be a three-layer fully connected neural network, with 196 input features corresponding to the EEG signals in the input layer, a first hidden layer with 3 neurons, a second hidden layer with 1 neuron, and finally two output neurons (calm/stressed) for binary classification. The weighted sum of each hidden neuron goes through a sigmoid activation function before being fed to the next layer.

The shallow feedforward ANN model was trained with the Adam optimizer and cross entropy loss function to penalize more heavily for misclassifying training examples and faster convergence. To avoid overfitting, 300 epochs of training were conducted. The learning rate was 0.01. The same neural network architecture (see Fig. 1) was used to evaluate the performance of different feature selections.

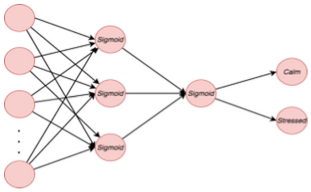

Fig. 1. The 3-layer ANN architecture selected.

2.3 Feature Selection with GA

To extend and improve on the approach in our first phase, we utilize a genetic algorithm for feature selection. Each individual in the population represented one subset of the available EEG features. Figure 2 shows how representing the inclusion/exclusion of the features as bit strings translates into a fully connected neural network.

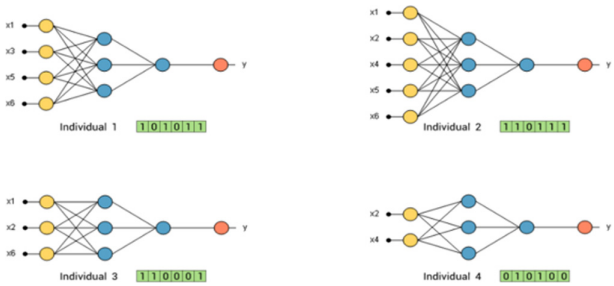

Fig. 2. ANN representation after feature selection with GA [11].

The population size of 20 was initialised randomly. We used an individual's average 5-fold testing accuracy as the measure of fitness, consistent with the performance measure in Sect. 2.6. The selection of the surviving individuals into the next generation was done with the proportional method – according to their relative fitness in the population.

Following selection, a new population was bred through crossover. Uniform crossover was used for this purpose. We chose uniform crossover because with so many available features, it's not feasible to manually pick justifiable crossover points. There are also no obvious contiguous regions in the EEG feature space. For additional exploration, the GA mutated the bit strings in the new generation with a small rate of 0.02 per attribute, giving an expected number of mutations of about 4. We also incorporated elitism which ensured that the population's best fitness is never decreasing and enabled the retention of good solutions. The proportion of elites to retain at each generation was 10%, which means the top 2 individuals were always directly put into the next generation, without mutation.

2.4 Sensitivity Analysis on Characteristic Patterns

One of the primary goals of this paper is to detect decision boundaries along input feature space, by identifying where there is a large rate of change in the output nodes with respect to the input nodes. It is theorized that where a small change in the value of an input causes the output to change from 0 to 1 or vice versa, it is likely to be the location of a decision boundary [7]. By searching for such decision boundaries in all input parameters, we produce simple rules to predict the output of the underlying ANN.

The fully-trained neural network is differentiated to find the gradients of the input features. For each output node (one for each class out of K), we calculated the gradient of input x for observation i. For each of the input features, a value z is found is where the absolute value of this gradient is largest. In Engelbrecht et al. [7], the decision boundary for each input feature is found by computing the gradient for every input pattern in the dataset, plotting a graph of the gradients and using a curve fitting algorithm to locate the peak gradient. The corresponding input value will then become the decision boundary. However, this is computationally expensive and thus, characteristic patterns are used to compress and represent the training set, as described in Gedeon et al. [5]. This reduces the number of gradients calculated.

Computing the gradient of each input for every pattern in the dataset was done over the characteristic patterns which represent the "typical" input pattern for each output class. Using the fully-trained ANN described in Sect. 2.2, each input was classified as *calm* or *stressed*. For each of these classes, a characteristic pattern (for example class k_1) was calculated as the arithmetical mean vector of all the relevant input patterns (all that the ANN classified as). For this EEG dataset, there are two characteristic patterns.

2.5 Rule Extraction

For each characteristic pattern, a gradient was calculated for each pair of output and input node. Only a set number of input variables with high gradients (in absolute value) were picked for each characteristic pattern to become rules. Since these will be the highest gradients, they are the most significant for determining the output and are most likely to be close to decision boundaries. Once a predetermined number of the highest gradients was selected, the corresponding inputs features become rules. A value on either side of each boundary was sampled to determine which class the ANN will predict if the input attribute were smaller/larger than the boundary value [7]. Thus, each characteristic pattern will have its own set of rules generated. In this paper, the number of rules extracted was 5. This allows the rules to have relatively high predictive power, while being interpretable. Too many rules would be hard for humans to make sense of.

When classifying a new input, it is first grouped into one of the characteristic patterns by Euclidean distance. Then the rules belonging to that characteristic pattern was run against the new input – giving a classification of *calm* or *stressed*.

For each characteristic pattern, the 5 rules were formed as the condition for an unseen input to be predicted as the class that isn't the characteristic class. We call these 5 rules the "rules against" the canonical class.

2.6 Performance Measure

As the provided EEG dataset is perfectly balanced, accuracy can be reliably used [8] as the evaluation measure. A train-test-validation split of the dataset was used to tune hyper parameters as well as give an unbiased evaluation of the chosen model. 20% of the data was used for the final evaluation, with the remaining 80% used to tune hyper parameters. Each neural network setting was evaluated with the average testing accuracy from the hold-out test sets during a 5-fold cross validation. Using rules extracted from the optimal ANN trained with all training data, predictions from these rules were compared to both the ANN outputs and the ground truth labels.

3 Results and Discussion

Overall, the ANN, rule extractions, and feature selection by GA all achieved a high level of performance, with the feature selection in phase two especially able to improve on previous results. We present the final results in Table 1.

Table 1. Final evaluation results – average final testing accuracy over 100 runs. The ANN results are not affected by the number of rules.

	Manual feature selection			GA feature selection		
	1 rule	5 rules	10 rules	1 rule	5 rules	10 rules
ANN against ground truth	–	93.8%	–	–	95.4%	–
Rules against ANN output	54.4%	85.0%	85.6%	57.1%	84.9%	90.3%
Rules against ground truth	54.5%	86.2%	87.5%	56.2%	86.3%	90.6%

3.1 Classification and Feature Selection Results

In the first phase of our study, the chosen 196-3-1-2 architecture with sigmoid activations achieved high cross-validation accuracy of 94.2%, and a final testing accuracy of 93.8% averaged over 100 runs. In the second phase, the fittest subset of features was selected according to the genetic algorithm. After 20 generations, 98 attributes out of the 210 available were retained. This is a very significant reduction in dimensionality by 53%. Using these features to train an ANN, we obtained 95.4% accuracy. This is an accuracy improvement of 1.6% compared to the testing performance of the ANN model without GA feature selection. Despite only using a small population of 20 and a low number of generations,

the GA was able to eliminate over half of the input features as redundancies, achieving better generalization.

No previous research used this identical dataset therefore we couldn't conduct a direct comparison. However, there is related research experimenting on the same task of stress detection. One such paper is Irani et al. [2]. For this experiment, no physical sensors were used and instead, computer vision techniques are utilized. Accuracies of 60% and 82% are obtained using RGB and thermal modalities, respectively. Compared to the higher of the two (82%), the results of this paper using EEG sensors and rule extraction does outperform it.

3.2 Rule Extraction Results

Using the default setting of 5 rules, the average testing accuracy of the rules when compared to the output of the phase one ANN reached 85.0%. After adding GA feature selection, the rule accuracy vs ANN prediction was 84.9%. This is a good level of accuracy considering the simplification from a continuous function-approximating ANN to a discrete one and much reduced set of simple rules. The high accuracy is in large part due to the effectiveness of the use of characteristic patterns. The characteristic patterns themselves are an encapsulation of the ANN's behavior.

The rules are useful mostly for classifying outliers, as only exceptional conditions will allow it to predict against its typical class. This is an explanation for the testing accuracies against ground truth labels of 86.2% (manual) and 86.3% (GA), which are higher than the rules' accuracies against the ANN output. As the number of rules extracted increases, the classification accuracy improves. As this sets a higher threshold for a prediction to be made against the default characteristic class, the predictive behaviour approaches that of the underlying ANN. However, there is a tradeoff between accuracy performance and interpretability. Having too many rules is difficult for humans to understand.

4 Conclusion and Future Work

This paper presented a shallow neural network built on EEG signals as a classifier for human stress. Combined with a GA based feature selection and sensitivity analysis-based rule extractions, a smaller, more powerful and more interpretable neural network achieved up to 86.3% accuracy. The ANN model prior to rule extractions reached 93.8% and 95.4% accuracy with manual and GA feature selection, respectively. Our work highlights EEG data as a key component in stress recognition research. In conjunction with a simple neural network model, a person's stress level can be reliably recognized. We also resolved the high dimensionality issue of EEG by adding a genetic algorithm to identify the most relevant features to be incorporated into the ANN.

By introducing rule extraction, we add evidence that neural networks can be explainable and hence, support its wider usage even in sensitive domains. Experts can be confident deploying neural networks to solve problems, with

rules as a "sense check" to provide an additional layer of assurance. Our paper further implements characteristic patterns for faster computation, but a possibility for future work would be extracting unconditional rules that do not rely on these characteristic patterns. This would require the sensitivity analysis to be performed over the entire training set. Such an approach could result in better decision boundaries and achieve a similar level of accuracy with fewer rules.

References

1. Rahman, J., Gedeon, T., Caldwell, S., Jones, R., Jin, Z.: Towards effective music therapy for mental health care using machine learning tools: human affective reasoning and music genres. J. Artif. Intell. Soft Comput. Res. **11**(1), 5–20 (2020)
2. Irani, R., Nasrollahi, K., Dhall, A., Moeslund, T., Gedeon, T.: Thermal superpixels for bimodal stress recognition. In: Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, Finland, pp. 1–6 (2016)
3. Lupien, S., Maheu, F., Tu, M., Fiocco, A., Schramek, T.: The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition. Brain Cogn. **65**(3), 209–237 (2007)
4. Saeed, S., Anwar, S., Khalid, H., Majid, M., Bagci, U.: EEG based classification of long-term stress using psychological labeling. Sensors (Basel, Switzerland) **20**(7), 1886 (2020)
5. Gedeon, T., Turner, H.: Explaining student grades predicted by a neural network. In: Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), Nagoya, Japan, vol. 1, pp. 609–612 (1993)
6. Hailesilassie, T.: Rule extraction algorithm for deep neural networks: a review. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **14**(7), 371–381 (2016)
7. Mira, J., Sánchez-Andrés, J.V. (eds.): IWANN 1999. LNCS, vol. 1607. Springer, Heidelberg (1999). https://doi.org/10.1007/BFb0100465
8. Chawla, N., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl. **6**, 1–6 (2004)
9. Erguzel, T., Ozekes, S., Tan, O., Gultekin, S.: Feature selection and classification of electroencephalographic signals: an artificial neural network and genetic algorithm based approach. Clin. EEG Neurosci. **46**, 321–326 (2014). https://doi.org/10.1177/1550059414523764
10. Babatunde, O., Armstrong, L., Leng, J., Diepeveen, D.: A genetic algorithm-based feature selection. Int. J. Electron. Commun. Comput. Eng. **5**, 889–905 (2014)
11. Gomez, F., Quesada, A., Lopez, R.: Genetic algorithms for feature selection. Neural Designer Data Science and Machine Learning Blog. https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection. Accessed 21 June 2021
12. Kalas, M., Momin, B.: Stress detection and reduction using EEG signals. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 471–475 (2016). https://doi.org/10.1109/ICEEOT.2016.7755604