Improving Casper Network using Genetic Algorithm on Hyperparameters Selection

Shanhe You
College of Engineering and Computer Science
The Australian National University
u6692394@anu.edu.au

Abstract. Cascade correlation algorithm [6] and its modification, the Casper (a Cascade Correlation neural network with RPROR optimizer [2]) has proven to be a useful structure in neural networks. A modified version of the Casper algorithm is examined in this paper, and further improvement will be analysed using a genetic algorithm for hyperparameters selection on the simplified Anger data set. The Anger dataset [1] used contains the pupillary movement data of the observers when they were reviewing real and posted anger emotion expression video. A standard genetic algorithm was examined to improve the hyperparameters setting of the modified Casper algorithm, on a relatively simple task of classifying real or posed anger video using the pupillary movement data of the observers in the Anger dataset, and the selection of hyperparameters including the three learning rates, their relationship and the number of epochs for training new neuron were optimized with the GA model. The Casper algorithm with GA achieves a higher accuracy of 88.92%, while the original setting of the modified Casper algorithm only achieves 85.25%.

Keywords: Genetic algorithm, Neural Network, Cascade Correlation Algorithm

1 Introduction

1.1 Motivation

1.1.1 The Casper algorithm

The original Casper algorithm has shown success in beating the original Cascor algorithm (Cascade Correlation algorithm) [6] on a complex task (the two Spirals benchmark problem) [2]. The Casper has shown potential on solving complex task, but its ability to solve simple task has rarely been mentioned or examined, this paper will examine the performance of the Casper algorithm on the simple classification task, and research possible improvement on its stability of performance, which has rarely been investigated before, further discover of its performance on small task can help investigate the potential of Casper algorithm.

1.1.2 The Genetic Algorithm

The genetic algorithm (GA) is a search algorithm developed based on the rules of natural selection and genetics [3]. It represents the idea of a way of intelligent exploitation of random search and has shown competitive for solving optimization problems [4].
The selection of the learning rate in the Casper algorithm is very important and will impact the result heavily, as it involves three different learning rate, three learning rates also represent its difficulty to optimized with manual experiments. A GA can be a good way to solve this problem and worth examination. In assumption, the standard GA should be able to determine satisfied hyperparameters for the Casper algorithm.

1.1.3 The Anger Dataset

Anger is one of the most important emotions of human, and it can be detected faster comparing to happy or upset faces, this is suggested to be part of the bias in human evolution for responding to threats faster [5]. But a fact was addressing that human cannot identify posed anger from genuine anger, while there is a difference in biological signal when human react to them.[6]. The success of using neural network for classification posted and genuine emotion using observers' pupillary data, can be helpful in effective HCI design to track user feedback [7], it will allow easier addressing and analyzing of the users' actual feedback on a very low cost comparing to manual analyzation.

1.2 Aims

This paper aims to develop a modified version of the Casper algorithm with the original RPROP optimizer and the idea of pool selection of neurons from the original Cascor algorithm [6], for achieving a more stable and compact Casper algorithm on solving relatively simple tasks.
Moreover, a genetic algorithm will be examined to optimize the selection of the hyperparameters of the modified Casper algorithm, including the selection of three learning rates, their relationship and the number of epochs. This will allow a comparison to evaluate the efficiency of using the GA algorithm on the Casper algorithm.
The Anger dataset will be used for evaluation of the models, where the model will be used to classify whether the video is posed or genuine based on the observers' pupillary movement data.

Overall, in this paper, an improved Casper algorithm will be developed, with hyperparameters selection using a genetic algorithm, the model aims to determine between real anger and posed anger and compare efficiency with a modified Casper without GA selection and a simple three-layer neural network.

1.3 Dataset

The Anger dataset used in this project is a simplified version of the original dataset, to examine Casper on a relatively simple task. The simplified dataset has 9 features as columns, and 400 rows, each row represents the pupillary movement data of one observer on the corresponding video, there are a total of 20 observers and 20 videos, which sum up to be 400 rows of data.
The 9 columns of features include the Observer, the video index, Mean, Std, Diff1, Diff2, PCAd1, PCAd2 and Label. The Observer and video index are simply index variable that indicates which observer and which video the data are referring to. The Mean and Std are the data of eye summary of eyeball varies, including movement and changes in sizes. The Diff1 and Diff2 are the records of differences between two eyeballs. The PCAd1 and PCAd2 are the results of an orthogonal linear transformation with the first and second principal component accordingly. The label simply defines the ground truth whether the anger in the video is posed or genuine.

1.4 Data Preprocessing

The features of Observer and video index are removed as redundant for this task, the Label is transformed into 0,1 values, where 0 for genuine and 1 for posed anger. The other 6 features are the data collection of observer's eye gaze data, including the change in eye size and movement and the difference between two eyeballs.
By plotting the box plot, the distribution of each feature can be identified clearly. A box plot of the data of 6 features of eye collection is shown in figure 1 below:

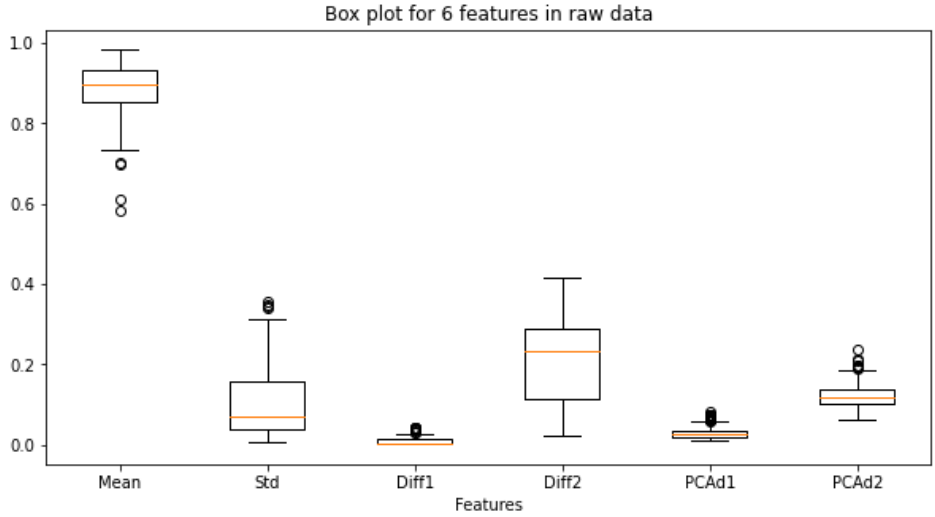
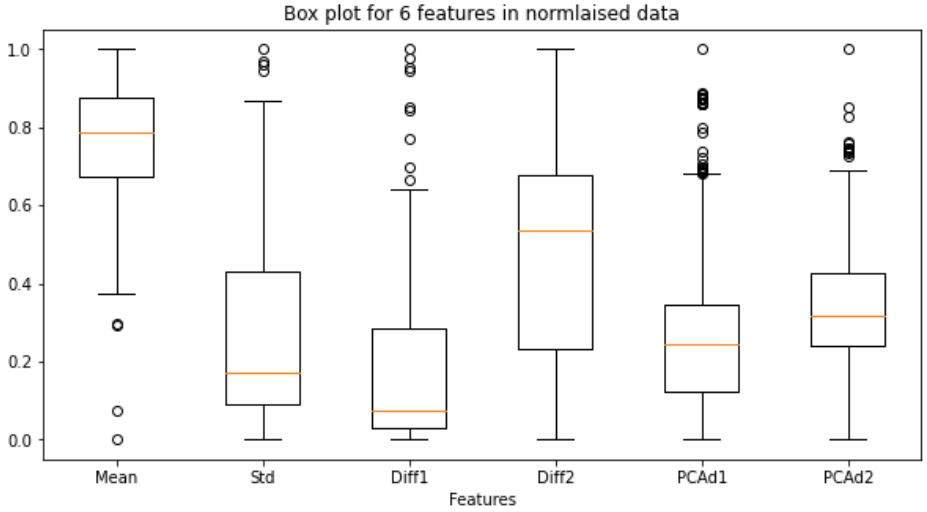


Figure 1. Box plot for raw data

It is clear that the raw data of these features distribute in a certain range, and the outliners are not too far from the region. The values of mean mostly distributed in the range (0.75, 0.98) with few outliners. The value of standard deviation distributes in the range (0.30, 0.008), with few outliners with a value up to 0.36. The value of Diff1 only distributes in a very small range (0.001, 0.044), where the differences are not obvious. The value of Diff2 distributes relatively uniformly compared to other features, in the range (0.002, 0.41) without outliners. Both of the PCAd1 and

PCAd2 have distribution in a small range, their range is (0.001,0.006), and (0.06,0.17) accordingly, both of this two feature has a relatively large number of outliners compare to other features, but the outliners are still very close to the main distribution, and thus, can treat as normal values in my opinion.

Summing up, the features have a tight distribution of data where the data distribution in a small range and the outliners are not too far from the main distribution. Thus, I suggest no removal of the outliners and normalize these data into the range [0,1] for increasing the differences between data, which helps in further training of the model. The box plot for normalized data shown next page in figure 2:



Fiagure2. Box plot for normalized data

We can observe that the data has a more discrete distribution, which is more beneficial for the network to classify comparing to the original data.
The equation of this normalization is:

$$x = \frac{x - \min(X)}{\max(X) - \min(X)}$$

where X is the data of the whole feature and $x \in X$.

2 Methods

2.1 The Cascade Correlation Algorithm

The Cascade Correlation algorithm [8] was proved to be a useful structure of the neural network, where the algorithm starts with a single hidden neuron, and adding new trained neurons into the network constantly until optimized (error decrease less than the setting ratio), the algorithm has two key ideas: The first is the cascade architecture [8], where hidden neurons are freezing and stop changing after added to the network. The second is the correlation maximization process of the new hidden neurons, where a process of maximizing the correlation between the new output and the residual error signal [8], which is maximizing the accuracy of the new hidden neurons. Figure 3 from the original paper [8] is shown below showing the structure of a Cascade Correlation neural network:

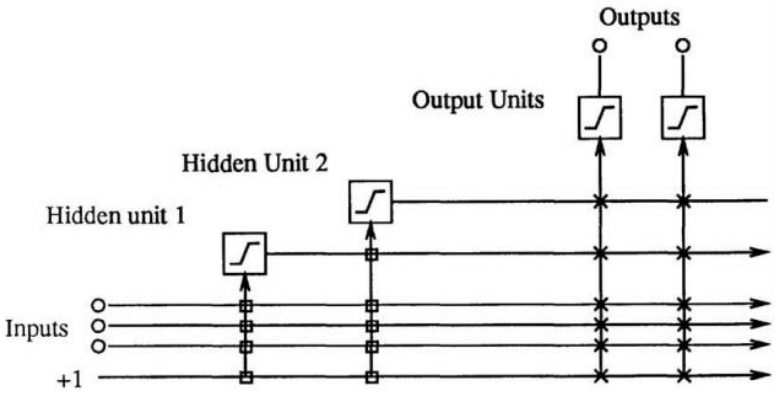


Figure 3: The Cascade architecture after two hidden neurons added. [8]

In the figure is the Cascade neural network structure with two neurons, where the boxed connections are frozen, X connections are trained repeatedly, meaning the parameters of old neurons are frozen.

2.2 The Casper Algorithm

The Casper different from the Cascade Correlation (Cascor), use a variation of RPROP to train the whole network and has shown to produce a more compact network that generalizes better than Cascor [2].

The Casper algorithm has similar architecture to the Cascade Correlation algorithm, where the network start from one single neuron and new hidden neurons are added into the network constantly. One of the main differences of the Casper algorithm is the usage of three different learning rate using different weights on a different region of the network. Assign the three learning rates to L1, L2 and L3, the figure 4 show a sample structure of the Casper architecture when the third hidden is added:

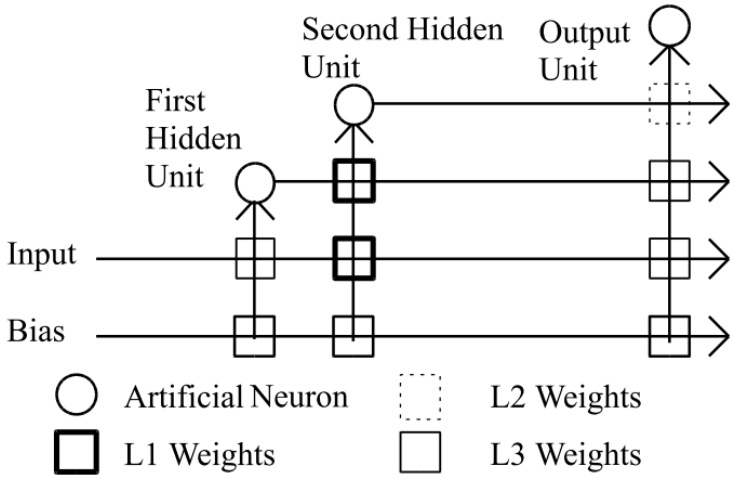


Figure 4: The Casper architecture when the second hidden neuron is added.[2]

The three-learning rates will have different weight and bias during the evaluation process:

> The L1 weights include all weight and bias connecting the new neuron from the old hidden neuron and inputs.
> The L2 includes all weight and bias connecting from the output of the new hidden neuron to the output layer.
> The L3 includes all the other weights and bias, meaning all weight connecting from previous hidden and input neurons.

Another difference is the removal of the weight freezing in the Cascor algorithm, wherein the Cascor algorithm, the weight and bias of old neurons are freezes, and only the weight of the new neuron will be updated when training and stop after the neuron is added to the hidden layer. In the Casper algorithm, the weight of three learning rate will change as the region vary, and the weight of old neurons will be updated together when the new neuron is trained, this avoids the situation of bad performance old neurons in the network and requires one or few new neurons to fix that.
By the removal of weight freezing, the Casper generate a more compact and stable result as lesser neurons are required, without the problem of using new neurons to fix old neurons.

2.3 The Modified Casper Algorithm

Two main modifications are done on the Casper algorithm for possible performance improvement.

In my implementation, the Simulated Annealing modification in the RPROP algorithm is removed, which was designed to solving the problem of getting a local minimum rather than the global minimum when training the new neuron, this is become a problem when solving complex tasks, while has limited improvement on solving the current classification task after testing, moreover, the other modification can also solve the problem of trapping in a local minimum. Thus, the Simulated Annealing modification was removed, and the original RPROP algorithm was used in the training of the modified Casper model.

The other modification is the usage of pool selection of the neurons, which was a method used in the Cascor algorithm, where a pool of new hidden neurons is generated and examined, only the best neuron will be selected and added to the

actual network. In my implementation, the size of the pool is ten neurons, meaning when a new hidden neuron needed to be added, ten neurons will be generated and evaluated, the parameters of the best neuron will be selected and applied to the new hidden neuron. This method not only solves the problem of getting into local minimum but also generate a more stable and compact performance of the network, as a smaller number of neurons are required with relatively high quality of new neurons.

2.4 The Genetic Algorithm

In my implementation, a standard genetic algorithm is used for determining the initial value of three learning rate, and the number of epochs for training new hidden neurons.
The genetic algorithm has similar principles of natural evolution, where a pool of random selection of values in the given range will make, which is called population, the selection of the values of the individuals in the population will be known as genes, the genes can be combined into the chromosome.
The fitness function which is used to determine how fit the individuals are is chosen to be the same as the testing accuracy of the model, which is a straightforward way to evaluate the setting of parameters.

The main genetic algorithm involves the selection phase, crossover phase and mutation phase:

The selection phase is to select the best individuals with high fitness and pass their genes (selection of values) into the next generation.

In the crossover phase, a crossover point is chosen randomly for each pair of parents, then the exchanging of the genes happens until the crossover point.

After the crossover phase, mutations are happening with a selected probability, where the individual mutated into values.

The selection, crossover and mutation phase are repeated until the population converged, and can only generate limited differences, or the maximum number of iterations is reached.

2.5 Detail Settings

For the GA algorithm, the model from the geneticalgorithm library is used with the following parameter settings:

> A maximum number of 15 iterations is selected, the algorithm usually converges within 10 iterations, and 15 provides a safe bound for possible exceeding. A population size of 10 is used, as each model will take a long time to train, a larger population will be too consuming, while 5 population sometimes cannot generate a good result after testing. The mutation probability, crossover probability and the parent's portion are kept as the default value, which are 0.1, 0.5, 0.3 accordingly. The elite ratio was set to 0 as a standard GA is implied. The max iteration without improving is set to 1, meaning the algorithm does not stop directly when the difference of population is lower than a certain threshold, it will iterate for one time, and stop if the differences are still less than the threshold.

The range of the learning rate Lr3 is [0.001, 0.5], the learning rate Lr2 and Lr1 are computed based on the Lr3, by multiple the Lr3 with the selected ratio, the range of the ratio of Lr2 and Lr1 are [2,20], [10,100] Accordingly, the number of epochs has the range [100,1000]. Only the learning rate Lr3 is continuous as float number, the ratio of Lr2 and Lr1 to Lr3 are integers, and the number of epochs is an integer.

After computing the optimized hyperparameters of the Casper model 10 times, averages are taken for the hyperparameters selection, the final examination setting of the L3, L2, L1 and number of epochs are 0.05, 0.91, 1.12 and 500 accordingly. Notice these values are rounded for easier usage, the original values are 0.05,0.90975, 1.12037, 535 accordingly, this small amount of rounding will not affect the result much and gives a more reasonable and accepted setting.

The loss function used in the model is CrossEntropyLoss, and the active function is the Sigmoid function.

3 Result

The data tested are the data from the simplified Anger dataset with pre-processing as described. The data were shuffled and randomly split with a ratio of 75% of the training set and 25% of the testing set. A validation set is not used in this task as it requires extra data usage, while the dataset is small, extra distribution of validation set will affect the date amount used in training and testing, and may lead to further decrease in the performance of the model. Overfitting will rarely happen on such a small dataset as well, thus, there is no need for a validation set.

My model of the modified Casper algorithm will stop adding new neurons after the new hidden neurons cannot improve the accuracy Furthermore when the model reach such states, the accuracy of the model will stay consistent or decrease if keep training, and the accuracy rarely increase. This, in my opinion, can be addressed as the model is well-trained and optimized.

All the result provided below were round to two decimal places.

The result of a simple Neural network with 10 hidden neurons is also used for comparing the performance. The simple neural network has training epochs of 5000 and the same data setting as the above models, the simple network will stop if the validating accuracy decrease and roll its parameters setting back to the optimized setting for best performance and avoid overfitting from happening. The average accuracy of the simple neural network is 77.08%, with a standard deviation of 9.80%.

The result of the original Casper algorithm was also concluded from 30 tests on the data, with the same condition of setting including the splitting of the data and learning for three sections. The model achieves an average accuracy of 86.32% and a standard deviation of 6.33%. The average number of neurons in the optimized network is 3.63, and the maximum number of hidden neurons is 8 and the minimum is 3.

The result of my modified version of the Casper algorithm was computed from 30 runs on the data, and the data were re-shuffled and split after each test. The model achieves overall averaging accuracy of 85.25% with a standard deviation of 5.52%. The average number of hidden neurons in the optimized network is 3.20, and the maximum number of hidden neurons is 7 and the minimum number of 2.

The result of my modified version of the Casper algorithm with hyperparameters (three learning rate and the number of epochs) selected by the genetic algorithm was computed from 30 runs on the data, the hyperparameters are selected and rounded by 10 runs as well, the detail number is provided in the section 2.5 Detail Settings. The model reaches the highest accuracy amount 4 models of 88.99% with a standard deviation of 4.09%, the average number of hidden neurons is 5.20, the maximum number of hidden neurons is 9, and the minimum number is 3.

A table is given below for the data comparison: (the Std refer to standard deviation)

	Simple NN	Original Casper	Modified Casper	Modified Casper with GA
Accuracy (%)	77.08	86.32	85.25	88.99
Std. (%)	9.80	6.33	5.52	4.09

Table1: Comparison of the accuracy and standard deviation

A graph showing the learning curve of the genetic algorithm is provided below:

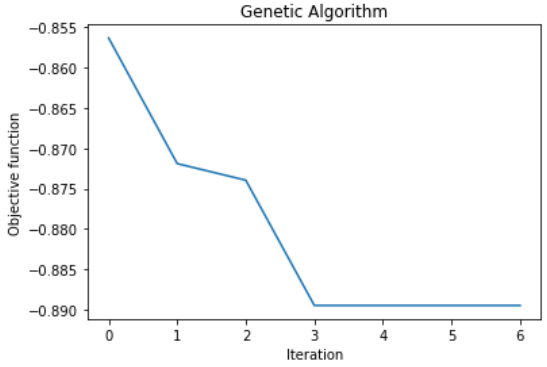


Figure 5: The learning curve of the genetic algorithm

The GA algorithm improves the accuracy of the model significantly in 3 iterations, and after that, the performance varies in a very limited range and can be treated as converging.

4 Evaluation and Discussion

4.1 Evaluation of the result

The result of the Simple Neural Network is undertaking on a three-layer neural network, with one input layer, one hidden layer and one output layer, there are 10 hidden neurons in the hidden layer, the learning rate of the model is 0.1 and the number of epochs is 500.

Both of the original Casper and the Modified Casper algorithm use three learning rate Lr3, Lr2, Lr1 of 0.05, 0.25, 1 accordingly with a number of epochs of 500.

The Casper with GA uses the learning rates of Lr3, Lr2, Lr1 of 0.05, 0.91, 1.1 accordingly with a number of epochs of 500.

From the results, it is clear that any of the Casper models perform much better than the simply Neural Network model, both in the accuracy and the stability of the performance.

There are no big differences in the model's accuracy and stability between the modified version of the Casper algorithm and the original one, the only difference is that the modified version of the Casper algorithm compiles overall slightly faster than the original one, which can be caused by the differences in the RPROP optimizer.

The modified Casper has a significant improvement in both accuracy and stability of performance comparing to the other two models, which shows the benefits of using GA in hyperparameters selection of the Casper algorithm.

4.2 Discussion

Summing up, the Casper model shows the potential of using physiological data (pupillary movement data in this task) to distinguish between posed anger and genuine anger successfully, and the Genetic Algorithm successfully improves the modified Casper algorithm both on the accuracy of the classification and the stability of the performance.
The modified Casper does not show a significant advantage above the original one with achieving slightly lower accuracy overall, but slightly higher stability. In my opinion, the decrease in accuracy can be possible due to the removal of the Simulated Annealing in RPROP optimizer, meaning the idea of pool neurons cannot solve the problem of converging into local minimum very well, and the Simulated Annealing in RPROP is still required even for a relatively small and simple task.
Although the genetic algorithm is very time consuming, the current setting of the algorithm requires approximately half an hour to compute, it is still worthy to do such computation, as it was higher the overall accuracy by around 3%, and increase the stability of the model furthermore.

Overall, the modification of the Casper algorithm does not show a good advantage above the original one, while both models can solve the classification of distinguishing real anger using physiological data, a standard genetic algorithm has shown significant improvement on the modified Casper model, and has shown its potential on complex parameters selection problems.

4.3 Future Work

Although the modified Casper does not generate an improvement as expected, the Casper algorithm has shown good potential on classification task using physiological data. The further experiment can be combining the Recurrent Neural Network (RNN) and the Casper algorithm, under such combination, the raw data set of the Anger can be further analysed as it is time-series data, which is one of the main advantages of using the RNN model as it will take time relationship into account. The RNN and its modification LSTM has shown good potential nowadays. The combination of the Cascor with RNN/LSTM has shown an advantage above the original RNN/LSTM model [9], it is worthy and interesting to investigate the possible improvements bring from the Casper to the existing RNN/LSTM models.

Another direction of the future work is investigating the genetic algorithm further on the Casper algorithm, due to the time limitation and the property of time-consuming genetic algorithm, only a very limited set of the genetic algorithm has been examined. A larger experiment with a larger population can improve the hyperparameter selection even more. An examination can also be done on using a genetic algorithm for hidden neurons training, as the structure of Casper that one hidden neuron is trained at a time, allow better and clearer observation of the performance of the genetic algorithm comparing to the normal neural network model.

5 Reference

[1] Chen, L., Gedeon, T., Hossain, M. Z., & Caldwell, S. (2017, November). Are you really angry? Detecting emotion veracity as a proposed tool for interaction. In Proceedings of the 29th Australian Conference on Computer-Human Interaction (pp. 412-416).
[2] Treadgold, N. K., & Gedeon, T. D. (1997, June). A cascade network algorithm employing progressive RPROP. In International Work-Conference on Artificial Neural Networks (pp. 733-742). Springer, Berlin, Heidelberg.
[3] Lucasius, C. B., & Kateman, G. (1993). Understanding and using genetic algorithms Part 1. Concepts, properties and context. Chemometrics and intelligent laboratory systems, 19(1), 1-33.
[4] Whitley, D. (1994). A genetic algorithm tutorial. Statistics and Computing, 4(2), 65-85.
[5] Mather, M., & Knight, M. R. (2006). Angry faces get noticed quickly: Threat detection is not impaired among older adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 61(1), P54-P57.
[6] Kim, C. J., & Chang, M. H. (2015, November). Actual emotion and false emotion classification by physiological signal. In 2015 8th International Conference on Signal Processing, Image Processing and Pattern Recognition (SIP) (pp. 21-24). IEEE.
[7] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., ... & Levinson, S. (2004, October). Bimodal HCI-related affects recognition. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 137-143).
[8] Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.
[9] Michielli, N., Acharya, U. R., & Molinari, F. (2019). Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Computers in biology and medicine, 106, 71-81.

