Effectiveness of Casper on EEG data compared to feed-forward networks using a variety of pre-processing techniques on detecting the stressed state an individual instantaneously

Galappaththi Chathura
Research School of Computer Science,
Australian National University
u6947345@anu.edu.au

Abstract. Traditionally stress detection techniques are unreliable as they don't use physiological responses. This study investigates the usefulness of EEG data in predicting whether an individual is stressed or calm at any given instance. The predictions are conducted using a modified version of Cascor known as Casper which shows potential when combined with effective pre-processing techniques; channel selection and feature selection, to reduce the dimensionality of the data. A method using variance and another using a genetic algorithm will be investigated to obtain the optimal channels and features. The Casper network performed better than a regular feed-forward network in some while performing worse in others however this could be due to ineffective feature selection as it performed better using the variation method than the genetic algorithm method.

Keywords: Casper, Cascor, EEG, Channel Selection, Feature Selection, Genetic Algorithms, Optimisation

1 Introduction

Since the start of the COVID-19 pandemic, conditions have remained uncertain with people losing jobs at an alarming rate and constant lock-downs across the globe. With all this uncertainty, it begs the question, what impact is this having on people's mental health? Results from polls show that a large number of adults are experiencing mental health issues as a direct result of the stress caused by this uncertainty (The Implications of COVID-19 for Mental Health and Substance Use, 2021). According to a paper written on using thermal super-pixels for bimodal stress recognition (n.d.), traditional methods are unable to detect stress instantaneously or continuously with those that can, being based on self-reporting which reduces accuracy. The same paper outlines that it is much more reliable to use physiological responses for detecting stress than any of these methods. As stress causes a change in electrical activity in the brain (Gaikwad and Paithane, 2017), Electroencephalogram (EEG) data would be an ideal source for this purpose.

There are many different machine learning techniques which could be used to predict stress based on the data from the EEG. One particularly powerful method is the Casper network. It is based on Cascor which proved to be effective for training neural networks much faster. Cascor starts with just the input and output layers then trains and adds a single neuron at a time whilst freezing all previously added neurons. This makes Cascor very fast as back propagation is essentially not used and the earlier epochs take less time due to the reduced number of hidden neurons (compared to a regular feed-forward neural network). The downfall of this method is that the early neurons are poor feature detectors and can lead to a larger network than a typical feed-forward neural network. Casper addresses this issue by allowing previous weights to be modified but at a much smaller rate. This means Casper has the benefit of having a reduced number of hidden neurons but without the disadvantage of having a larger network.

This report will implement the Casper algorithm as outlined in the paper by Treadgold and Gedeon (n.d.) and apply pre-processing techniques, namely channel and feature selection to determine its affect on the accuracy of the model. This is because it has been shown that irrelevant features reduce performance (Kohavi and John, 1997) in addition to increasing computational complexity. The reduced set of channels (ie. an electrode) and the features of each of these channels (eg. mean, min, max, etc), will be selected by using the variance of the data and by using a genetic algorithm. These reduced features and channels will then be fed to a Casper network as well as a regular feed-forward neural network and compared against each other to determine the effectiveness of Casper. In addition to this, each of these will be compared with a baseline Casper and feed-forward network with minimal pre-processing.

2 Method

2.1 Feed-forward neural network implementation

As a baseline comparison, a regular feed-forward neural network has been implemented. An arbitrary learning rate of 0.2 was used for this network as it resulted in a good balance of the network accuracies and the training time. The network consisted of a single hidden layer with the number of hidden neurons set to half the number of input neurons. This was done to force the network to obtain a reduced representation of the features allowing it to generalize better. In addition to this, it has the effect of reducing the training time as less computations are required.

2.2 Casper implementation

The Casper Algorithm (Treadgold and Gedeon, n.d.) was implemented using the methods outlined in the paper by Treadgold and Gedeon (n.d.). As the paper by Treadgold and Gedeon, n.d., did not specify the hyper-parameters, the learning rate used for the feed-forward neural network was used for L1 as it allows for a better and more fair comparison. In addition to this, arbitrary learning rates of 0.005 for L2 and 0.001 for L3 was used. This was picked through trial and error for the same reason the learning rate was chosen in Section 2.1; it strikes a balance between the accuracies and training time. The paper by Galappaththi (2021) which implemented the same Casper algorithm, used a P value of 4 and obtained excellent results. Due to this, the same P value will be used by Casper network discussed in this paper.

2.3 Basic pre-processing

Basic pre-processing was done on all datasets used in this experiment. The EEG data consists of 211 features and a single binary label. One of these features is the subject number and as it is not a determining factor of whether an individual is stressed or calm, that is you can swap around the subject numbers without the outcome changing, it was excluded from all experiments conducted. This is further supported by the results of the paper written by Irani et al. (n.d.) which states physiological responses are the most reliable for stress detection and as the subject number is not an indicator of a physiological response, it can simply be removed.

2.4 Pre-processing using variance

As discussed above, irrelevant features can result in a decrease in the performance of a network. With this in mind, it may be possible to reduce the number of channels and their features without having an impact on the performance, potentially even increasing it. A paper investigating feature extraction and channel selection of EEG signals for seizure diagnosis (2021), used the variation across channels to select three which best fit their purpose. The same concept can be applied to the dataset of this experiment by calculating the variation of the average value of each channel (ie. the mean feature of each channel) then picking the n most varied channels where n is found experimentally to produce the best results. As per the paper written by Galappaththi (2021), n was set to 2 as it produced the best results.

The features used for each channel can be reduced by using a similar concept. To accomplish this, all data must first be grouped by the features, that is group the data by the mean, min, max, skw, etc then the variance of the data in each group can be calculated to determine the m features with the largest variance. As per the same paper referenced above, m was set to 13 as it produced the best results (in this paper k is used for the number of features).

2.5 Pre-processing using genetic algorithms

Following on from the goal in Section 2.4, a genetic algorithm could be used to select the channels and their respective features rather than simply using the variance. This is because using the variance may remove some channels and features which are crucial for predicting the outcome accurately.

2.5.1 Encoding

To use genetic algorithms, the problem needs to be encoded as a chromosome which can then go through the process of natural selection to arrive at an optimal solution. For this purpose, each channel and feature can be represented as a gene with a binary allele which determines whether that channel/feature is included in the dataset used to train a network. This means that each chromosome will have the structure shown in Figure 1.

F3	O1	T8	P7	AF4	AF3	O2	F7	FC6	F8	T7	FC5	F4	P8	min
mean_first_diff	fuzzy	hurst	sum	apen	rms	hjorth	skw	iqr	mean_second_diff	std	var	mean	max	

Figure 1. The encoding of a chromosome which represents the optimisation problem

Each generation will consist of fifty chromosomes and each will be used to produce a reduced dataset which will then be passed through to a Casper network where it will train on this dataset and output a testing accuracy. K-fold cross validation will be used to run the Casper network k times where the testing accuracy from each run is used to obtain an average fitness which is more representative of true performance of that genotype (a visualisation of this can be seen in Figure 2). This is necessary due to the stochastic nature of the Casper network initialisation which can cause the same chromosome to have varying accuracies. In addition to this, it is common place for experiments with less data (as is the case with this experiment) because it shows general performance.

In an ideal world, the k value will be large however due to computational limitations, k was chosen to be 15 as it resulted in each generation taking roughly two minutes to evaluate the average fitness of each generation. This was enough of a balance between obtaining a general testing accuracy and time taken to train the underlying Casper networks.

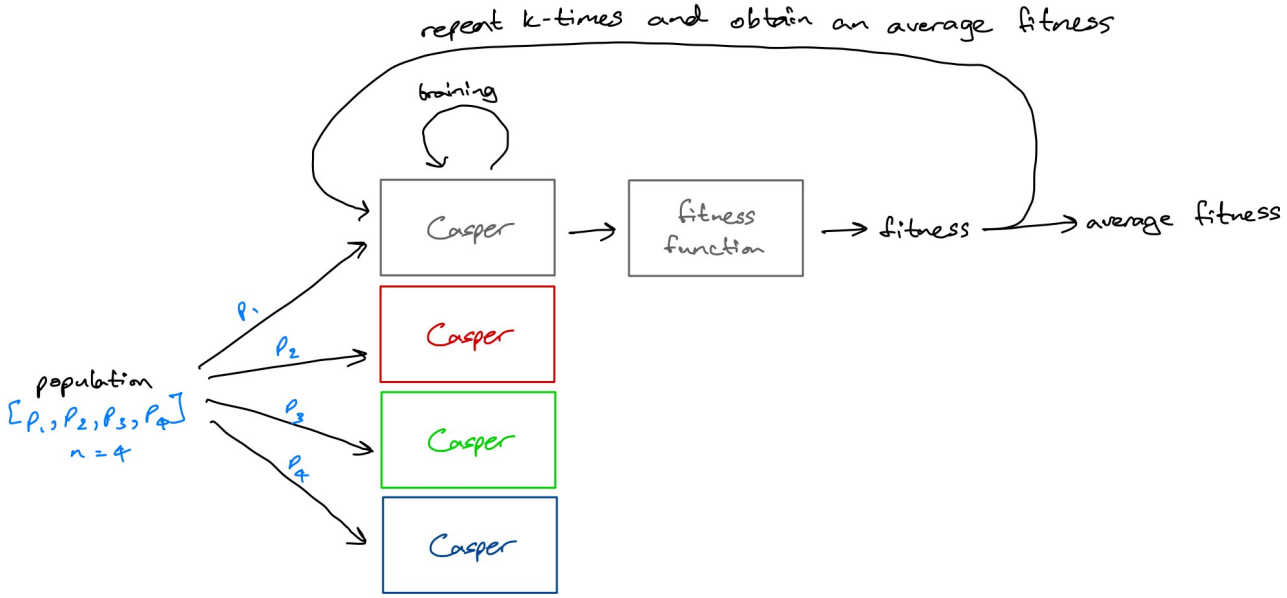

Figure 2. The process of obtaining the average fitness for each genotype of each generation

2.5.2 Fitness function

As the purpose of this network is to obtain a feature set which can generalise well, the minimum testing accuracy obtained across the k Casper networks is used as the fitness function. This will mean the algorithm will optimise to increase the minimum performance and with enough time this will theoretically yield to a population where at their worst a respectable accuracy is still obtained. Using the average was also considered but due to the point mentioned before, the minimum was used instead.

2.5.3 Generation progression

The algorithm is initialised with a random population of size fifty. This was chosen as a paper which utilised genetic algorithms to optimise the horizontal axis of a wind turbine (Pourrajabian, Dehghan and Rahgozar, 2021) found that a smaller population combined with a larger number of generations, speeds up the convergence rate. Due to this, a population size of fifty (rather than 100 or 200) with a larger number of generations such as the 83 used in this paper, would be optimal for this purpose. Once the population has been initialised, the average fitness value of the generation will be calculated as explained in Section 2.5.1.

Once the average fitness value has been calculated for each genotype, a subset of the population will be discarded based on its average fitness value and the pass through rate (arbitrarily chosen to be 0.8 for this case). The remaining population will reproduce to create children using uniform crossover to replace those that were discarded. Uniform crossover (where each gene is selected randomly from one of the parents) was used as it means that there can be more of a variation across the chromosomes compared to other forms of crossover. In addition to this, a paper written by Pourrajabian, Dehghan and Rahgozar (2021) has discovered that it "could improve the convergence rate of the binary genetic algorithm".

For each child that is produced, it has a chance of obtaining a mutation based on the mutation rate (arbitrarily chosen to be 0.01 for this purpose) which will further increase the variation across the chromosomes which can lead to discovering superior genes faster. This is especially useful after a few generations where the gene pool is very similar where the mutation can discover a gene which helps overcome the diminishing performance of future generations.

3 Results and Discussion

3.1 Casper vs feed-forward neural network with basic pre-processing

A Casper network and a feed-forward neural network was trained with the basic pre-processing discussed in Section 2.3, that is almost all the data except for the subject number was used for predictions. The results of both networks can be seen in Table 1 below

Table 1. Performance of Casper and a feed-forward neural network (FFNN) with 100 epochs and a P value of 4 over 150 trials (10 repeats of different 15-fold cross validation trials) using all channel and features

	Network	Mean	Std	Min	Max
Hidden Neurons	Casper	3.97	0.27	1	4
	FFNN	105	0	105	105
Training Accuracy	Casper	76.26	3.42	54.47	91.04
	FFNN	54.30	11.06	50.00	100.00
Testing Accuracy	Casper	48.70	12.16	22.22	80.00
	FFNN	40.10	7.34	30.00	60.00

As expected both networks perform poorly due to the presence of many irrelevant features however, Casper performs better than the feed-forward neural network in general. This is most likely due to Casper not having a limit on the number of neurons allowed therefore it can continue training and extracting features to get a better score. As stated previously, the number of hidden neurons in the feed-forward network is half the number of inputs. As the data consists of 14 channels each with 15 features, it means the network had 210 inputs therefore 105 hidden neurons in its layer.

3.2 Casper vs feed-forward neural network with channel and features selection via variance

As explained in Section 2.4, a different dataset was then used to evaluate the variance across the channels and the features to reduce the dataset down to 2 channels and 13 features. The reduced dataset was then passed into Casper and a feed-forward neural networks, the performance of which can be seen in Table 2 below.

Table 2. Performance of Casper and a feed-forward neural network (FFNN) with 100 epochs and a P value of 4 using 2 channels each with 13 features run over 150 trials (10 repeats of different 15-fold cross validation trials)

	Network	Mean	Std	Min	Max
Hidden Neurons	Casper	3.99	0.11	3	4
	FFNN	13	0	13	13
Training Accuracy	Casper	71.90	4.49	60.00	80.07
	FFNN	80.60	3.16	70.15	91.79
Testing Accuracy	Casper	63.44	13.67	22.22	100.00
	FFNN	72.49	13.55	33.33	100.00

Compared to Casper with basic pre-processing, the data above shows a decrease of 6% in the training accuracy and an increase of 30% in the testing accuracy. An increase of 48% in the training accuracy and 81% in the testing accuracy

can be seen for the feed-forward neural network when compared to the same network with basic pre-processing. There is a performance increase in both models which further shows that irrelevant features are detrimental to a network's performance. It can also be seen that the feed-forward neural network performs much better than Casper with this reduced dataset as it has a testing accuracy that is 14% better than Casper. This suggests that Casper may not be suitable for this dataset however it could be due to many factors. Firstly, Casper has more hyper-parameters than the feed-forward neural network therefore it is possible that the current hyper-parameters (P value, L1 rate, L2 rate and L3 rate) are sub-optimal. In addition to this, even though both networks have the same number of epochs for training, Casper's accuracy drops significantly after a new neuron is added therefore it is at a slight disadvantage specially if training is stopped directly after a new neuron has been added.

3.3 Casper vs feed-forward neural network with channel and features selection via a genetic algorithm

As explained in Section 2.5, a genetic algorithm was designed find the optimal channels and features which results in a high testing accuracy. The channels and features determined by this algorithm was then used to reduce yet another dataset which was passed into a Casper and a feed-forward neural network and its results can be seen in Table 3 below.

Table 3. Performance of Casper and a feed-forward neural network (FFNN) with 100 epochs and a P value of 4 using 5 channels each with 8 features run over 150 trials (10 repeats of different 15-fold cross validation trials)

	Network	Mean	Std	Min	Max
Hidden Neurons	Casper	4	0	4	4
	FFNN	20	0	20	20
Training Accuracy	Casper	76.19	3.46	68.15	93.28
	FFNN	55.22	12.03	50.00	100.00
Testing Accuracy	Casper	48.84	11.94	22.22	80.00
	FFNN	40.61	8.30	22.22	70.00

Compared to Casper with channel and feature selection using variance, the data above shows an increase of 6% in the training accuracy but a decrease of 23% in the testing accuracy. In contrast, the feed-forward neural network saw decrease of 31% in the training accuracy and 44% in the testing accuracy. In fact, the accuracies obtained using the genetic algorithm only performs slightly better than the version which only performed basic pre-processing. As Section 3.2 has shown that a reduced dataset can lead to a significant increase in performance, this suggests that the genetic algorithm is ill-suited for this purpose. However, there is no reason why in theory, the genetic algorithm cannot match or surpass the results obtained in Section 3.2 as the algorithm would eventually obtain the same set of channels and features used in that section. The poor performance of the genetic algorithm could be due to a range of factors. Firstly, the number of generations used may not be enough to converge to a good result or the number of times, k, the Casper network was re-trained may not be enough to obtain an accurate representation of the general performance of that genotype. Secondly, it may be that the pass through rate (ie. the number of genotypes which continues on to the next generation), is too high or too low which means not enough of a variance is introduced or too much of a variance is introduced which is negatively impacting the gene pool. Finally, the fitness function may be a very poor indicator of the performance of that genotype as it only looks at the testing accuracy and disregards the training accuracy and other factors.

It is important to note here that Casper outperforms the feed-forward neural network. This suggests that the Casper network could prove to be effective given the right hyper-parameters and a dataset which has all the irrelevant features removed.

5 Conclusion and Future work

Although Casper performed worse using the reduced dataset via variance, it out-performed the feed-forward neural network in all other runs. However, the accuracy of Casper is far from optimal as the highest testing accuracy obtained is 63% which is almost equivalent to guessing. As the network does show improvements with some pre-processing, it has potential and could potentially be very effective given an optimally reduced dataset.

There are many ways in which this model could be improved. Firstly, the Casper model itself could be improved by utilising a different loss function such as Adam which appears to be one of the best modern optimisers. Using an activation and loss function which is more prevalent in modern applications may also yeild better results which could

potentially allow this network to obtain a much higher testing accuracy whilst maintaining a relatively high training accuracy. In addition to this, another genetic algorithm could be used to determine the hyper-parameters of Casper, the P value and the L1, L2 and L3 learning rates.

Secondly, it is clear that the genetic algorithm used to obtain a reduced dataset has not been configured correctly, that is the current hyper-parameters are not well suited for its purpose. To improve this, the network could be run for longer (ie. increase the number of generations) and with a higher mutation probability to allow for more of a variance which could lead to better performing genes. Following on, the pass through rate could also be decreased to allow for more children with varying genes which could increase the speed of convergence. It is important to note however, that this may also reduce performance as the children may end up disrupting the existing gene pool. A larger k value could also be used as it would result in a fitness value which is more representative of the genotype. In addition to this, the fitness function could also be changed to use the median value rather than the minimum and average as it could better represent the performance of the genotype. A feed-forward neural network may work better to obtain the reduced dataset which can then be fed into a Casper network to potentially obtain better results. Finally, a different crossover method may also improve performance as uniform crossover may be changing too many of the good genes.

References

1. Ein Shoka, A., Alkinani, M., El-Sherbeny, A., El-Sayed, A. and Dessouky, M., 2021. *Automated seizure diagnosis system based on feature extraction and channel selection using EEG signals*.
2. Gaikwad, P. and Paithane, A., 2017. *Novel approach for stress recognition using EEG signal by SVM classifier*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/8282611> [Accessed 27 April 2021].
3. Irani, R., Nasrollahi, K., Dhall, A., Moeslund, T. and Gedeon, T., n.d. *Thermal Super-Pixels for Bimodal Stress Recognition*.
4. KFF. 2021. *The Implications of COVID-19 for Mental Health and Substance Use*. [online] Available at: <https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/> [Accessed 27 April 2021].
5. Khoo, S. and Gedeon, T., n.d. *Generalisation Performance vs. Architecture Variations in Constructive Cascade Networks*.
6. Kohavi, R. and John, G., 1997. *Wrappers for feature subset selection*.
7. Rahman, J., Gedeon, T., Caldwell, S., Jones, R. and Jin, Z., 2020. *TOWARDS EFFECTIVE MUSIC THERAPY FOR MENTAL HEALTH CARE USING MACHINE LEARNING TOOLS: HUMAN AFFECTIVE REASONING AND MUSIC GENRES*.
8. Treadgold, N. and Gedeon, T., n.d. *A Cascade Network Algorithm Employing Progressive RPROP*.
9. Pourrajabian, A., Dehghan, M. and Rahgozar, S., 2021. *Genetic algorithms for the design and optimization of horizontal axis wind turbine (HAWT) blades: A continuous approach or a binary one?*. Elsevier.
10. Galappaththi, C., 2021. *Application of Casper to detect the stressed state of an individual instantaneously using EEG data and appropriate pre-processing techniques*. The Australian National University.