Alcoholic Subject Detection: EEG Classification using Casper Algorithm and Genetic Algorithm for Hyperparameter Selection

Wai Lok Lam

Research school of Computer Science
Australian National University
Canberra ACT 2600 Australia
u6768030@anu.edu.au

Abstract. Alcoholism is a common disorder which accounts for more than 5% of all deaths in the world, and often under-diagnosed as it relies on self-report, using other objective data might making the screening process easier. In this paper, we used the UCI EEG dataset for alcoholism, which contains the EEG data of control subjects and subjects that is diagnosed with alcoholism, which we then perform classification task and determine if the subject is alcoholic or not. We implemented the Autoencoder for feature extraction, genetic algorithm for hyperparameter optimization, with Casper algorithm, which is a modified Cascade correlation algorithm, to then compare the classification of both within-subject and cross-subject settings to other researcher's result. We found that the Casper network performs relatively well when it comes to within-subject settings with mean accuracy of 89%, but failed to produce outstanding result in cross-subject settings, with mean accuracy of 70%.

Keywords: Autoencoder, Genetic Algorithm, Neural network, Deep Learning, Cascade, Casper, EEG

1 Introduction

Alcoholism is a common disorder with prevalence of more than 8%, individual with alcoholism is often also diagnosis with other mental disorder concurrently [8]. It accounts for 5.3% of all deaths in the world [17] and related to other social issues. American Psychological Association (APA), the association that refers to the most in diagnosis of mental disorders, refers alcoholism as Alcohol Use Disorder in their diagnostic manual DSM-5 [18], which the diagnosis relies heavily on the individual's self report. This leads to alcoholism more often being under-diagnosed, as the reliance on self-report might not be effective as some patients would cover up such information [8],1there are attempts by researchers to use other objective method for the diagnosis of the Alcohol Use Disorder [9].

Electroencephalogram (EEG), which measured the potential difference caused by the current generated by neurons in our brain [1, 2], has been found extremely useful when it comes to diagnosis of certain disease and medical condition such as brain damage [3], epilepsy [1] and brain tumor [1], and disorders such as sleep disorder. along with the rising popularity of neural network and deep learning, classification and prediction using EEG data has been found to be very effective [5].

EEG data is extremely challenging to process and perform classification task due to it being a times series data, and the data measured by physical device, which subject to all sorts of human error and measurement error, including artifacts, which can be described as changes in the recording of the signal due to outer (unwanted) factors, including those arising from the equipment, the environment, the subject and misoperation [6], which might make the task much more difficult.

As an attempt to improve the performance of the classification task, in this paper, we implemented the Casper algorithm, proposed by Treadgold and Gedeon [7], we experiment the effectiveness of using the Casper algorithm, combine with the use of autoencoder for feature extraction and genetic algorithm for hyperparameters optimization on the UCI EEG Alcoholism dataset [13]. We then performed within-subject testing and cross-subject testing using the model, and compared the result with previous result published by different researcher.

2 Methods

2.1 Dataset

The dataset we used in this paper is from the UCI Machine Learning Repository, originally owned by Neurodynamics Laboratory of State University of New York Health Center [13]. The dataset consist the data of 11,057 trails of EEG

data, collected thought 122 subjects with 77 subjects with alcoholism and 45 subjects without alcoholism [9, 10], and around 120 trails has been performed for each subjects, where different stimulus (one of five picture in the *Snodgrass and Vanderwart picture set* [12]) were presented to different subject (and same subject in different trail) [13]. Each trail last one second, for each trail, 64 electrodes is placed on the subject and recorded the activity at 256 Hertz, which makes it extremely difficult to process as each trail would have 16,384 features (excluding other labels) [13].

Each trail is also paired with a subject id (1 - 122), the stimulus presented during the trial (labeled by 1 – 5) and the alcoholism label (labeled by 0 or 1). As each entry is labeled by subject id, we used the subject id for splitting dataset when performing 11-fold-cross-subject-validation. In all case, the subject id of all entries is removed before training, as it does not contain useful information and leaving the column during training might lead to the model relies it for classification.

2.2 Data Exploration and aim of this paper

The aim is to use the EEG channel data and the stimulus label to classify if the subject is diagnosed with alcoholism [10]. We performed data exploration to see if there is any interest relationship between field that can help us in choosing input features, we can see from the dataset that the stimulus presented during the trail is represented in 1 – 5, we can see that the stimulus used during the trail is not likely to be randomly selected, with more than 5500 times stimulus 1 is shown but only <100 for stimulus 5 (Fig. 1). As the stimulus is recorded a single column, we added 5 columns as one hot representation of the 5 stimulus used.

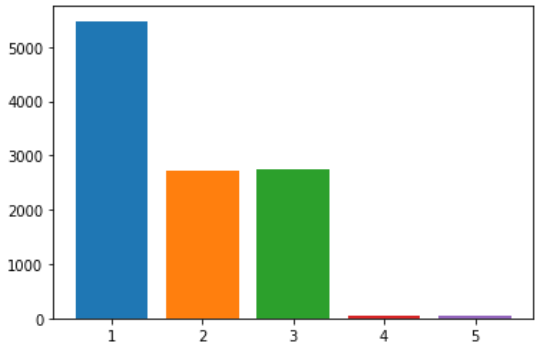

Fig. 1. The frequency of different stimuli used

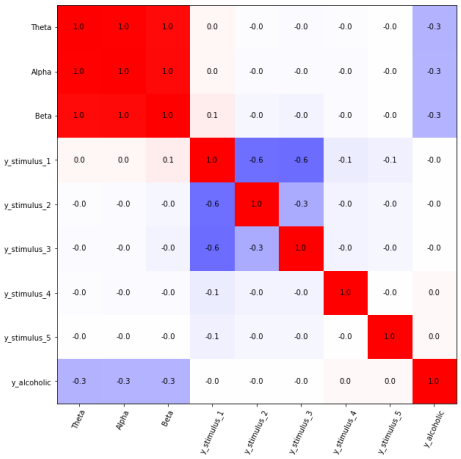

Fig. 2. Pearson's correlation between the mean value of each frequency band and labels

We investigated the correlation of different attributes of the data, measured by Pearson's correlation, we tried to inspect the correlation between each three frequency bands of signals (alpha (8 to 13 hertz), beta (13 to 30 hertz) and theta (4 to 7 hertz) and the alcoholic label, we took the mean value from each band and plotted the correlation matrix (Fig. 2). We can see that all three frequency band are equally correlated to if the subject is alcoholic or not (if we only consider the mean value of those channel).

2.3 Further preprocessing, Preprocessing method selection and Features selection

For further preprocessing, to determine what data should be used for the later training, we separated 10% of the subject for preprocessing method selection, the best hyper-parameter for the model, for determining both cross-subject and within-subject preprocessing methods, we used 10-fold-cross-validation and tested different method of preparing the data for training using Casper neural network. Only 10% of the subject has been used because we wanted to avoid the possibility of the preprocessing/hyper-parameter/model biased towards the specific dataset, picking 10% of the subject out and never uses it in later training and testing stage greatly reduces the change of it being biased.

Two different preprocessing dataset and further preprocessing method were chosen in this paper for within-subject testing and cross-subject testing.

Preprocessing Method for Within-subject Test. A preprocessing version of the original UCI dataset were used. Provided by Research school of Computer Science at the Australian National University, the preprocessing dataset consist 197 columns, 192 columns consist of the mean value of 64 channels * 3 frequency band per channel (i.e. alpha 1 – alpha 64, beta 1 – beta 64, theta 1 – theta 64), trail number, subject id (1 - 122), the stimulus presented (1 – 5) and the alcoholism label (0 or 1).

Preprocessing Method for Cross-subject Test. A preprocessed dataset used for cross-subject Test is based on a preprocessed version of the original dataset, encoded from raw EEG data into EEG images by Yao, Plested and Gedeon [12] using method based on work by Bashivan et al. [19], and provided by of the Research school of Computer Science at the Australian National University . The prepossessed dataset consist of 11,057 entries, representing 11,057 trails, each entries consist of a 32 * 32 * 3 colour, where each colour represents one frequency band (i.e. theta, alpha and beta signals is encoded to 32 * 32 image of red, green, blue receptively).

Remark 1. Both preprocessed dataset is based on the same exact original dataset, but preprocessed and provided by the Research school of Computer Science of Australian National University for us to use for this paper.

The dataset is extremely difficult to train, after flattening the dataset into a 11,057 rows * 3072 (32 * 32 *3) features dataset and train it using neural network (with Casper algorithm), we found that the neural network was not able to train efficiently (the loss is not reducing/stuck during training) and the overall result is very poor. Because of this, we tried to use other method such as doing 3D pooling on the data to preserve the spatial aspect of the dataset, however, the result seem to be even more worse than directly using 3072 features.

We then employed a autoencoder to capture the features of the dataset. Autoencoder is a principal component analysis technique, by using a neural network that consist of a bottleneck in the center to "compressed" the data and training the neural network to reproduce the input data [20]. Upon further testing using the separated model selection (10%) dataset, we used a autoencoder with 3 hidden layer, with the bottleneck layer consist of only 256 features (Fig. 3), and trained it using batch size of 512, learning rate of 0.001 and 25 epochs. After training the autoencoder and saving it, we then use it to encode (refers to the encoder part in Fig. 3) the 3072 features flattered dataset to 256 features for training in the neural network.

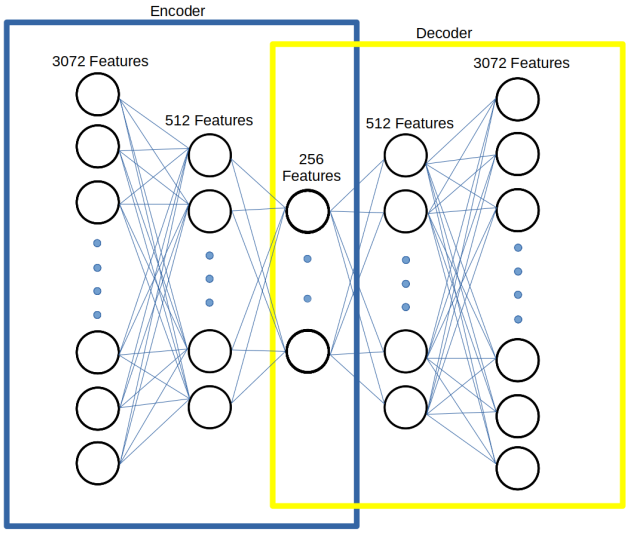

Fig. 3. The topology of the Autoencoder used for preprocessing

2.2 Casper algorithm

Cascade correlation (Cascor) algorithm [14], is an method that starts with zero neuron (input features directly connected to the output layer) and while freezing other neuron's weight, add one neuron to the model that takes previous neurons and input features as input and output to the final output layer of the whole model and train both the output layer and the new neuron, and repeat this process (of adding new neuron) until some conditions are met. The Cascade correlation algorithm has proven to have the capability to solve complex classification problem [15], and weight freezing means that back-propagation is not needed [7], making the algorithm much cheaper compared to other traditional method. However, the cascade algorithm suffers heavily from issue such the size of the overall network is unnecessary large due to it freezing the weight of previously trained neuron [7].

The Casper algorithm, proposed by Treadgold and Gedeon in 1997 [7], is an algorithm based on the Cascor algorithm, it uses the same structure as the Cascor algorithm, which is to add a neuron to the network and train the neuron, but unlike Cascor, the Casper algorithm does not freeze the weight and bias of other existing neurons, but to drastically decrease the learning rate of other existing neurons, it uses 3 different learning rate, L1, L2 and L3 for three types of neurons (Fig. 4), where L1 is significantly larger than L2 and L2 is slightly larger than L3. L1 rate is used for the weight of the newly added neuron, which is relatively high, letting the new neuron to learn minimize the error at a faster rate. L2 rate is used for the weight between the newly added neuron to the output neuron. Lastly, the L3 rate is used for all the other weights, including the weight between the input features and the old neurons.

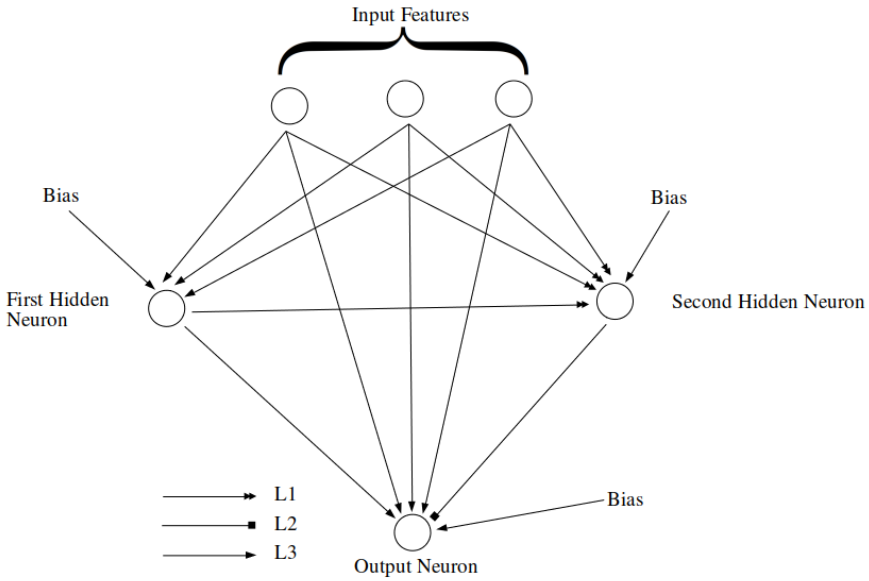

Fig. 4. The topology of Casper algorithm after adding the second neuron [7]

Casper algorithm uses resilience back-propagation (Rprop) [15] for gradient descent using the learning rate specification mentioned above along with weight decay with simulated annealing (SA), which is:

$$\frac{\delta E}{\delta w_{ij}} = \frac{\delta E}{\delta w_{ij}} - K \cdot sign\left(w_{ij}\right) \cdot w_{ij} \cdot 2^{\left(-0.01 \cdot E\right)} [7]$$

Where E meant the number of epoch that has been completed since the introduction of the last neuron, and K is a custom hyper-parameter, it reduces the gradient of the network along with the training, and resetting each time a new neuron is added to the network [7].

Additionally, for every *set amount* of reduction of loss after the introduction of a new neuron, the algorithm then add a new neuron to the network, if the network failed to reduce the loss by a *set amount* in $15+P*N$ epochs, where P is a custom parameter and N is the total number of neurons in the network, the whole network will halt [7].

2.3 Casper topology tuning

After performing the testing on the separated 10% of the subject's data, we found that increasing the weight of the neurons connected to the output layer allows the model to have overall lower loss rate compared to the original Casper algorithm, for simplicity, we change the learning rate of all weights and bias connected to the output neuron to *L2* (Fig. 5).

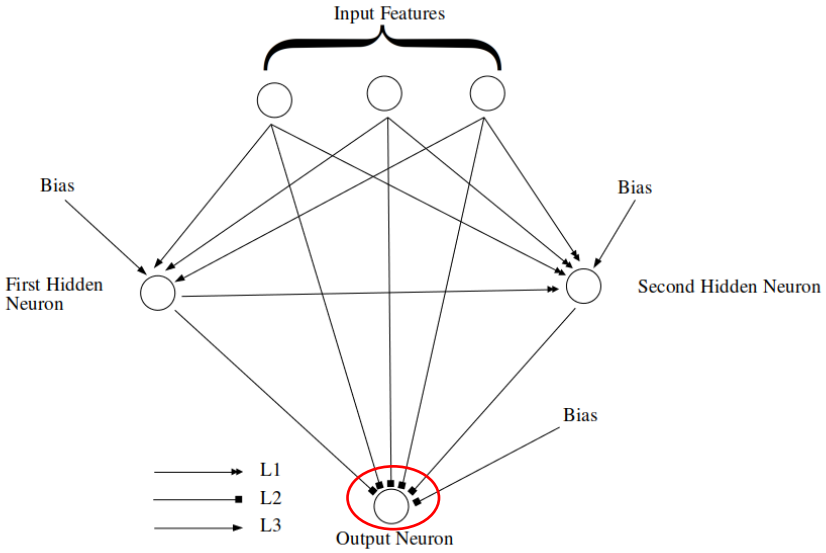

Fig. 5. The topology of Modified Casper algorithm after adding the second neuron (Difference from Fig. 4 is highlighted by the circle)

2.4 Genetic Algorithm & Casper hyperparameter optimization using Genetic Algorithm

Genetic algorithms, invented by John Holland, is an algorithm that imitate the process of natural selection, and perform "selection", "crossover" and "mutation" to select the best solution [21].

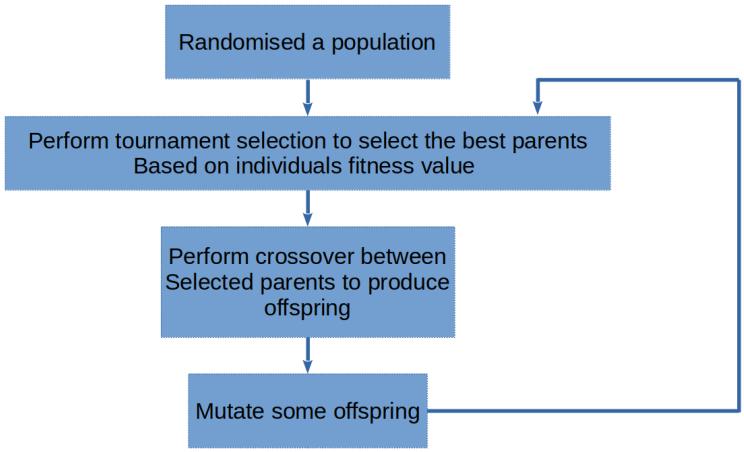

Fig. 6. The flowchart for genetic algorithm [21]

For the case in this paper, we will use Genetic algorithm to decide the best hyperparameter of the Casper algorithm for the classification task. We will use tournament selection (n = 3) as selection method, which perform a "tournament" between randomly chosen *n* individuals and the individual with higher fitness will be selected to be parent (each individual can enter a tournament unlimited times), and once all parent are selected (parent amount = population size), we then perform crossover [21]. For this task, five different hyperparameters of Casper is the "attributes" of each individual, which includes the (1) cutoff (maximum training accuracy before stopping); (2) the *K* parameter; (3) Required loss between each time set; (4) the initial constant in the time set formula; (5) maximum number of neuron. We evaluate the fitness based on a 2-fold-cross-validation using the separated validation subjects' data (using the hyperparameters), and perform uniform crossover between parents with crossover rate of 0.8 and mutate rate of 0.2 (0.2 chance of the individual mutating, and further 0.2 chance for each one of five parameters to mutate).

2.5 Result of Casper Hyperparameter optimization using Genetic Algorithm

We performed Genetic Algorithm with population size of 50 and 200 generations, the result of the selected hyperparameter is presented in the following table.

Table 1. The hyper-parameters used for with-subject and cross-subject settings

Parameter	Within-subject	Cross-subject
L1	0.2	0.2
L2	0.01	0.01
L3	0.005	0.005
K	0.0039	0.0077
P	50	50
cutoff	0.053	0.087
Batch size	Full batch	Full batch
Time set Formula	$55 + P * N$	$49 + P * N$
Required loss between each time set	0.014	0.012
Max number of neuron	14	7

2.6 Testing methodology

After finalizing the hyper-parameter mentioned above, we then used two different method to test the performance of the neural network. We will use the testing data presented by Li et al. [10] and Yao and Gedeon [12] as baseline to compare our model against it, both in within-subject test and cross-subject test. The paper uses 7:1:2 for within-subject test and 11-fold-cross-subject-validation for cross-subject, we will, however use 10-fold-cross-validation for within-subject testing as we found that result of single result of train-validation-test split is highly dependent on the split, and is not indicative of the performance of the overall model performance. And for cross-subject test, we used the same method as Li et al. [10].

3 Results and Discussion

The result were compared with other methods tested or used in Li et al. [10] or Yao and Gedeon [12]. The code is entirely written in Python, the Casper and the autoencoder part is written using PyTorch, and the genetic algorithm is written using the DEAP library [22]. The model uses codes in the documentation of the DEAP library and sample codes provided by Australian National University, it is trained using NVIDIA GTX1060, AMD RYZEN 5900x, 32 GB RAM using Ubuntu 20.04 environment.

3.1 Within-subject Testing

Table 2. Casper Within-Subject test result using 10-fold-cross-validation (2 decimal places)

Method	Mean	Median	S.D.
Time (seconds)	147.24	123.43	79.41
Total epoch	5529.8	6034	831.24
Number of neurons	13	14	1.63
Test accuracy	**0.89**	0.89	0.98
Final Loss	0.12	0.12	0.028
Sensitivity	0.85	0.85	0.019

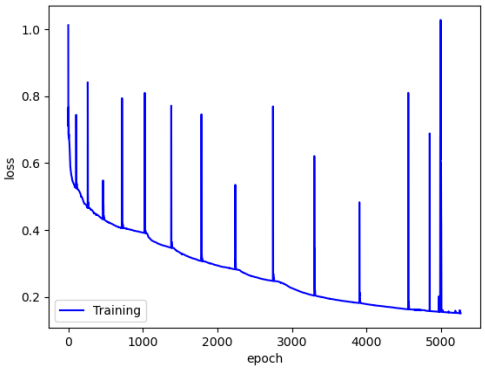

Fig. 7. Typical training loss-epoch graph for within-subject training

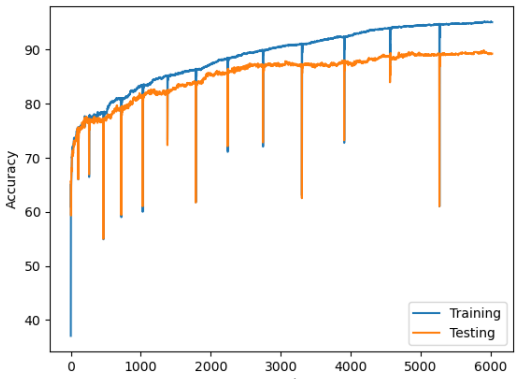

Fig. 8. Typical training accuracy vs test accuracy for within subject testing

For Within-subject test, like mentioned in above, we uses 10-fold-cross-validation evaluation the performance of the model on within-subject testing. We can see from Fig. 7 and Fig. 8 that the loss spikes up, and the accuracy drops drastically every time a new neuron is added, which is expected, and because the structure of the Casper algorithm, with high learning rate for the new neuron, the loss quickly drops to and then drops below the previous loss (vice versa for the accuracy). And we can see from table 2 that the test accuracy is very stable between splits with the standard deviation of 0.98.

3.2 Cross-subject Testing

Table 3. Casper Cross-Subject test result based on 11-fold-cross-subject-validation (2 decimal places)

Method	Mean	Median	S.D.
Time (second)	12.62	10.50	4.02
Total epoch	304.55	250	93.42
Number of neurons	1.27	1.0	0.47
Test accuracy	**0.70**	0.72	11.80
Final Loss	0.54	0.54	0.015
Sensitivity	0.63	0.64	0.21

For cross-subject test, we ran 11-fold-cross-validation to test the performance of the model, the result is presented in Table 2. We can see from the table that the model performs have much worse results compared to within-subject test. With the mean accuracy of 0.7 and a standard deviation of 11.8, we can see that the performance of the cross-subject settings is not consistent. And we can also see the number of neurons is rather low, what meant that the model cannot reach a certain loss increase during the training of the neuron (This potentially indicate that the model is not able to "train", or the training ends during the first neuron).

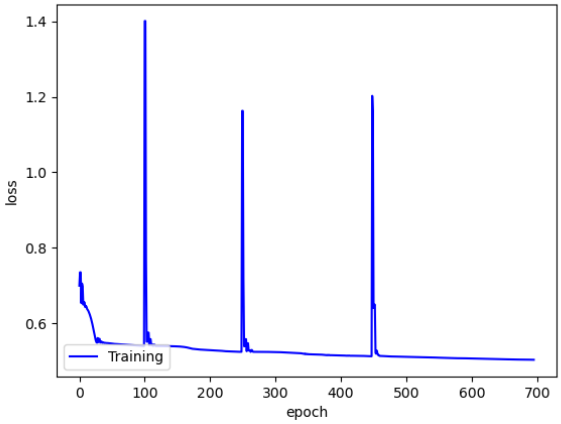

Fig. 9. Typical training loss-epoch graph for cross-subject training

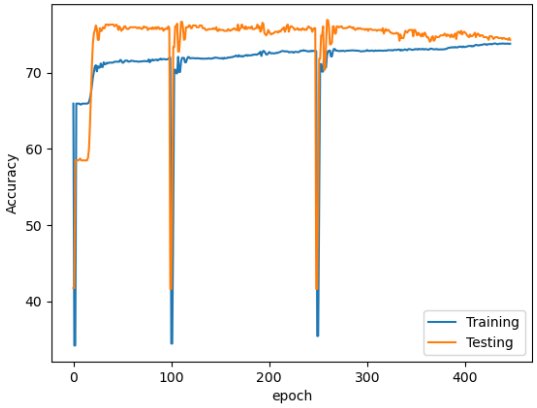

Fig. 10. Typical train-test accuracy graph for cross-subject training

We can see from [Fig.9] and [Fig. 10] that the training loss and the testing accuracy is not increasing during the training, which might indicate that the model is unable to learning though back-propagation, or the unable to pass a local minima.

3.3 Results compared to Li et al. [10] and Yao and Gedeon [12]

Table 4. Sorted Within-subject classification accuracy of the UCI Alcoholism dataset

Method	Within-subject accuracy
GP-SyncNet [10]	0.923
SyncNet [10]	0.918
Normal Image-wise Autoencoders [12]	0.917
Shared weight Image-wise Autoencoders [12]	0.897
Casper w/ genetic algorithm	**0.890**
EEGNET [19]	0.878
Normal Channel-wise Autoencoders [12]	0.864
Shared weight Channel-wise Autoencoders [12]	0.858
MC-DCNN [20]	0.840
DE [16]	0.821
PSD [17]	0.816
rEEG [18]	0.702

Table 5. Sorted Cross-subject classification accuracy of the UCI Alcoholism dataset

Method	Cross-subject accuracy
Normal Image-wise Autoencoders [12]	0.756
Shared weight Image-wise Autoencoders [12]	0.740
Normal Channel-wise Autoencoders [12]	0.731
GP-SyncNet [10]	0.723
Shared weight Channel-wise Autoencoders [12]	0.713
SyncNet [10]	0.705
Casper w/ genetic algorithm	**0.695**
EEGNET [19]	0.672
DE [16]	0.622
rEEG [18]	0.614
PSD [17]	0.605
MC-DCNN [20]	0.300

We can from Table 4, 5 that the Casper algorithm performs fairly well in within-subject settings with the mean score of 0.89, which is comparable to many other framework/model for EEG classification. In cross-subject settings however, we can see that Casper does not outperforms most of the other methods.

4 Conclusion and Limitation

Casper algorithm proposed by Treadgold and Gedeon [7] were implemented in this paper, together with genetic algorithm for hyperparameter optimization, we used the Casper algorithm to train neural network for classification on the EEG dataset to classify if a subject is an alcoholic. We tested our results in within-subject testing and cross-subject testing against the result published by researchers [10, 12] [Table 4]. We found that the Casper algorithm performs relatively good in within-subject testing and were extremely consistent with standard deviation of 0.98 in accuracy, the mean accuracy ranked 5 out of 12 different methods implemented in [10] and [12].

However, for Cross-subject validation, Casper seems to have mediocre performance compared to other results [Table 5], which the mean accuracy ranked 7 out of 12 different methods. We can see from the training-testing accuracy graph that (Fig. 9) and the loss-epoch graph (Fig. 10), the test accuracy and loss was capped in very early stage of training, and training seems to have no effect on the final accuracy on the model, we suspect that this might be due to insufficient work taken in the feature selection (only using autoencoder is not enough), or because of the limitation of the Casper algorithm.

This paper demonstrated that there are potential for using Casper for solving complicated EEG classification task, especially for the task that is designed for the purpose of within-subject testing or screening. And further testing and training using new data is needed for the model to be adopted in real life settings, which is often cross-subject settings.

5 Limitation & Future Work

There are several limitation of this paper. Firstly, the preprocessing of the EEG data is insufficient, for the within-subject settings, we can clearly see that there is no need for complicated feature selection process to achieve good result, however, in the cross-subject testing, we can see that the Casper algorithm is unable to train based on the features encoded by the autoencoder, we suspect this might be due the spacial features of the EEG data is loss when we flattened the dataset for encoding, for future work targeting this limitation, ways to capture spacial data for EEG task should be used.

Secondly, there are computational constrain during the use of genetic algorithm for model selection, as there are essentially more than 20 hyperparameters (considering different L1, L2, L3 combination for different layer) and different topology settings for the Casper network, in additional to the fact that evaluating the fitness of the model needs actual training and testing to be done on those hyperparameter each time. Exploring few hundreds generation with population size of a hundred can be extremely computationally expensive (and we speculate the process might take years). Although we suspect it is not possible, future work should investigate whether we can use some kind of shortcut or heuristic to determine the potential performance of those hyperparameter without actually need to train-test it.

Future work should be done on continuing the exploration of the possibility of using Cascade Correlation Network or the Casper network to perform classification task on physiological data, and experimenting on combining different preprocessing technique and training methods.

References

1. Schomer, D.L. & Lopes da Silva, F. 2010, Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Wolters Kluwer Health, Philadelphia.
2. Kirschstein, T. & Köhling, R. 2009, "What is the Source of the EEG?", Clinical EEG and neuroscience, vol. 40, no. 3, pp. 146-149.
3. Watanabe, K., Hayakawa, F. and Okumura, A., 1999. Neonatal EEG: a powerful tool in the assessment of brain damage in preterm infants. Brain and Development, 21(6), pp.361-372.
4. Siddiqui, M.M., Rahman, S., Saeed, S.H. and Banodia, A., 2013. EEG signals play major role to diagnose sleep disorder. International Journal of Electronics and Computer Science Engineering (IJECSE), 2(2), pp.503-505.
5. Craik, A., He, Y. and Contreras-Vidal, J.L., 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of neural engineering, 16(3), p.031001.
6. Klass, D.W., 1995. The continuing challenge of artifacts in the EEG. American Journal of EEG Technology, 35(4), pp.239-269.

7. Treadgold, N.K. and 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.
8. Enoch, M.A. and Goldman, D., 2002. Problem drinking and alcoholism: diagnosis and treatment. American family physician, 65(3), p.441.
9. Mumtaz, W., Vuong, P.L., Malik, A.S. and Abd Rashid, R.B., 2018. A review on EEG-based methods for screening and diagnosing alcohol use disorder. Cognitive neurodynamics, 12(2), pp.141-156.
10. Li, Y., Dzirasa, K., Carin, L., Carlson, D.E.: Targeting EEG/LFP synchrony with neural nets. In: Advances in Neural Information Processing Systems, pp. 4623–4633 (2017)
11. Sykacek, P., Roberts, S.J.: Adaptive classification by variational Kalman filtering. In: Advances in Neural Information Processing Systems, pp. 753–760 (2003)
12. Yao, Y., Plested, J. and Gedeon, T., 2018, December. Deep feature learning and visualization for EEG recording using autoencoders. In International Conference on Neural Information Processing (pp. 554-566). Springer, Cham.
13. EEG Database Data Set. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [Online], Available: https://archive.ics.uci.edu/ml/datasets/eeg+database, October 13, 1999.
14. Fahlman, S.E. and Lebiere, C., 1990. The cascade-correlation learning architecture. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.
15. Riedmiller, M. and Braun, H., 1993, March. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In IEEE international conference on neural networks (pp. 586-591). IEEE.
16. W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development, 2015.
17. World Health Organization, 2018. Global status report on alcohol and health 2018: Executive summary (No. WHO/MSD/MSB/18.2). World Health Organization.
18. AMERICAN PSYCHIATRIC ASSOCIATION. (2013). Diagnostic and statistical manual of mental disorders: DSM-5. Arlington, VA, American Psychiatric Association.
19. Bashivan, P., Rish, I., Yeasin, M. and Codella, N., 2015. Learning representations from EEG with deep recurrent-convolutional neural networks. arXiv preprint arXiv:1511.06448.
20. Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE journal, 37(2), pp.233-243.
21. Mitchell, M., 1998. An introduction to genetic algorithms. MIT press.
22. Fortin, F.A., De Rainville, F.M., Gardner, M.A.G., Parizeau, M. and Gagné, C., 2012. DEAP: Evolutionary algorithms made easy. The Journal of Machine Learning Research, 13(1), pp.2171-2175.