Genetic Algorithms and Network Pruning in Improving Deceit Detection Using Neural Networks and Pupillary Responses

Thao Pham

College of Engineering and Computer Science,
The Australian National University, Australia
u7205329@anu.edu.au

Abstract. Recent advancement in technology has allowed many occupations to adopt the hybrid working model and younger generations to grow up with social media and virtual identities. Inevitably, the spreading of false information has become increasingly prevalent. This study aimed to study the possibility of recognising dishonesty through a recording, using pupillary features trained on a neural network. We found that a baseline neural network predicted accurately 48.4%, lower than the chance level. As we suspected the problem of overfitting may be present, we looked into simplifying the model design through Genetics Algorithm and Network Pruning. Feature selection has helped us identify the three most important factors to the classifier: the amount of variation of the pupillary sizes throughout the viewing, the average changes from one period to another (given by first and second differences), and details regarding the peak occurrences. We also motivated the use of Network Pruning based on distinctive analysis to reduce the hidden layer complexity while maintains the neural network performance. We found that by eliminating 80% and 90% of the hidden neurons for the ANN and ANN + GA model, both models' performance increased to 56.3% and 53.5%, respectively.

Keywords: Machine Learning, Neural Networks (NN), Subjective Beliefs, Pupillary Responses, Network Pruning, Unit Pruning, Deceit Detection, Distinctiveness Analysis, Feature Extraction, Genetic Algorithms, Evolutionary Computation

1 Introduction

Given the rise of computational power in handling complex data, scientists have tapped into the wide possibilities of machine learning and artificial intelligence. We have seen intelligent systems classifying objects, recognizing voices, and understanding semantics. However, at the very core of it all is human emotions, internal thoughts and feelings, and the desire to stay connected. As a species that thrives when living in communities, we have learned to decipher others' body language to communicate better or to detect deception, which from an evolutionary standpoint, to survive better.

Culturally, the eyes have been thought of as a window to one's soul as it is reflective of one's emotions. Scientifically, researchers have also looked into the eyes, or more precisely, the changes in pupillary features over time, as indications of one's emotional or psychological state. Indeed, one's pupillary responses can be used to understand how the individual receives and reacts to certain information. Research done by Partala & Surakka (2003) has shown that one's pupils dilate when interacting with sounds of significant emotional responses (baby laughing or people fighting) and remains unchanged with neutral sounds. Another recent study carried out by Prochazkova, E. et al. has demonstrated that synchronization of pupil dilation between two communicating partners is a sign of trust (2018).

Linking this idea back to our current digital age, the abundance of availability of organisational and functional software has made it possible for many occupations to work remotely from home. This significant increase in online communication and decrease in direct physical contact raises the need to investigate human's ability in understanding other's virtual behaviours as well as how one's pupils adapt to lie detection in the new environment.

Newly-released research led by Zhu et. al (2018) has found that an artificial neural network (ANN) can be trained to detect falsehood from a recording using observers' pupillary responses throughout the video with an accuracy of 58.3%. Interestingly, this accuracy is significantly higher than that of the observers' verbal judgement on whether or not there existed untruthfulness in the said video, which was 50% (ibid.). It could be reasoned that the physiological responses are instinctive and thus corrupted by one's personal biases. This result motivates the need for more research into understanding pupillary responses and their reflection into one's unconscious biases.

The goal of this paper is to explore the most optimal subset of pupillary features in detecting one's subjective belief and whether it would improve the model performance by incorporating Genetics Algorithm into the baseline ANN. This paper also motivates the use of Network Pruning to find a more efficient structure for the ANN with equivalent performance.

2 Experimental Design

In this paper, we will use the dataset collected and produced by Zhu et al. in the mentioned paper. The experiment was conducted in two different stages: presenter recording and observer experiment. The first stage involved four presenters. Each was given four different topics. Two of which the presenter was led to believe to be true, while for the remaining two topics, they were told to be false information. The presenter then proceeded to present these materials to a camera. The second stage involved 23 new participants. Each had the task to watch the 16 recorded videos. As mentioned, the physiological responses of each observer were recorded throughout the session. At the end of each video, they were asked to determine whether or not the presenter believed in the material they just delivered. In this paper, we are not interested in the observers' verbal conclusions. Note that the effect of the sequence of videos being shown was considered as the playing order was randomised for each participant.

Finally, the data was organised to a final of 368 patterns. Each pattern represents an observer's physiological responses to each of the recorded videos. The signals included Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Skin Temperature (ST), and Pupil Dilation (PD). We will only look into Pupil Dilation and its ability to predict presenters' subjective beliefs. The reason why we are particularly interested in pupillary responses is, with working and socialising environments being slowly moved to virtual, human eyes are functioning in a very different environment as they did just a century ago. Therefore, we would like to investigate solely pupil dilations and their potential in detecting dishonesty.

The raw data relating to PD are sequences of each observer's pupillary sizes for both left and right eyes while watching each video. Missing data points due to occasional eye blinks were interpolated using cubic spine by the owner of the data to ensure cohesion across all patterns. Finally, this is a binary class problem with target values 0 for the presenter had his/her belief manipulated and thus has doubts about the material they presented (175 data points) and 1 for non-manipulated beliefs (193 data). Note that the distribution of these two classes is fairly balanced (47.6% and 52.4%, respectively). Thus, it is safe to the testing accuracy for evaluating model performance.

3 Methodology

3.1 Pre-processing

Due to the sequential nature of the data, Zhu had further extracted and processed the raw data into the following time-domain features:

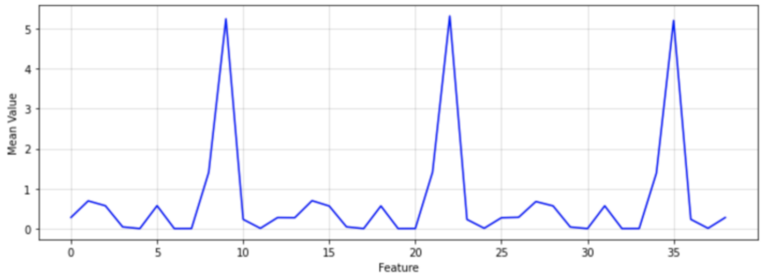
1. minimum pupillary size
2. maximum
3. mean
4. standard deviation
5. variance
6. root mean square
7. means of the absolute values of the first differences
8. means of the absolute values of the second differences

The author then performed further signal processing to generate five more features:

9. number of peak occurrences for VLP signals
10. number of peak occurrences for LP signals
11. amplitude of peak occurrences for VLP signals
12. amplitude of peak occurrences for LP signals
13. ratio of peak occurrences between VLP to LP signals

This procedure was done for the pupillary sizes for both left and right eyes, as well as an averaged value (mean of left and right values) which gives a final of 39 features for each training sample. As for our part, we inspect the data by taking the mean value across all features and plot it in Figure 1. The difference in ranges between each feature is evident as expected due to different natures of the features such as the number of peak occurrences will take on integer values while ratio strictly lays between 0 and 1. This suggests that a min-max scaler (0-1) is needed for all signals to eliminate the difference in scale and any unstable behaviour it may introduce. Apart from that, we choose not to further process the data. We only extract the described features relating to Pupillary Responses and proceed with our experiments.

Fig. 1. Mean values of 39 Pupillary Reponses features.



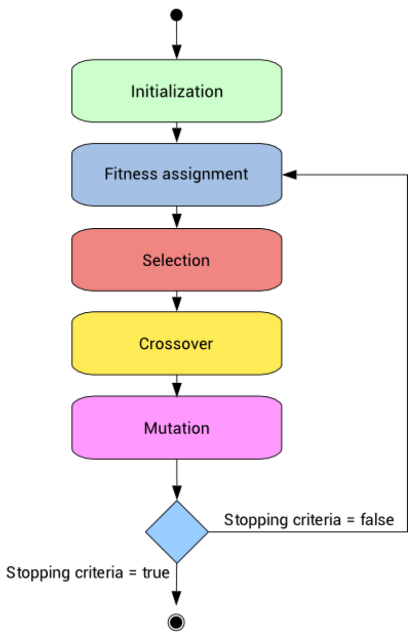
3.2 Classification / Baseline Neural Network

A three-layer ANN was employed to predict the presenter's belief based on the changes in pupillary features of the observers. The baseline ANN contained an input layer of 39 neurons (for 39 features), a hidden layer of 100 units with the sigmoidal activation function, and an output layer of two neurons for final classification. The reason why this specific architecture was chosen is that the exact model design was used in Zhu's paper. Using these parameters is beneficial for our goal to access whether feature extraction and network pruning can improve the accuracy achieved by said paper. The model was trained with the Adam Optimiser and back-propagated by the CrossEntropy Loss Function. We found that the combination of learning rate = 0.0013 and 700 epochs gives the best model generalisation.

Classification results were cross-validated using the leave-one-participant-out method by repeatedly setting all data from one participant to be a holdout set (16 patterns for each validation set) and the remaining data (352) patterns to be the training set. The final results were then averaged and reported.

3.3 Feature Selection

Given our main goal of finding the best subset of pupillary features in distinguishing subjective beliefs, we will apply the technique of Genetic Algorithms (GA) for feature selection. Explained by Zhu et al. (2018), the accuracy obtained using biometric features extracted by GA surpasses that of the features subset obtained by other selection methods. It is a reasonable finding as GA mimics the evolutionary mechanism of genetics – the cyclic trial-and-error and organisms design based on natural selection – and uses this to find optimal solutions to a problem. In our case, GA helps find a combination of pupillary dilation features that produces the best testing accuracy. This paper follows closely the GA procedure demonstrated by Gomez et al. (2020).



Define Representation & Initialisation: Firstly, we initialise the F0 generation as a set of N neural networks, which we have set N to be 390, 10 times the original number of features. Each network is similar to an individual in a population where its DNA is a binary sequence of length 39 (for 39 features) where 1 is equivalent to the feature is used for that model and 0 otherwise. The baseline model in which all pupillary dilation features are selected is also included for proper comparison.

Evaluate Fitness Values: Each network will then be trained with the hyperparameter settings similar to that of the baseline model. After training, we rank the models based on their testing accuracies. Note that this ranking system is ordered in an ascending fashion. The higher the testing accuracy, the higher is the rank of that model. The fitness value is then assigned for each model as follows where k is the selection pressure.

$$f_i = r_i \times k$$

The selection pressure is a hyperparameter that ranges from 1 to 2. If it is very close to 2, the discrepancy between the models with lower (Ls) and higher (Hs) accuracy becomes higher and thus will strongly increase the chance of Hs being chosen for reproduction.

Fig. 2. Genetics Algorithm overall procedure. From "Genetic Algorithms for Feature Selection" by Gomez et al. (2018).

Selection: We then employ the selection operator to choose individuals (neural network models) to participate in reproduction. This is done by first assigning a probability to be selected to each individual proportional to their fitness value. Note that two models with the same fitness values will share the same chance of having their genes passed down to the next generation. Following this, we will select N/2 individuals based on this probability distribution. Elitism is also used to ensure the fittest individuals will appear in the new population. We implement elitism by selecting the top 2% population. Half of this group will go straight to the new generation while the remaining half will go through the recombination procedure (bypassed the selection process).

Crossover & Mutation: After selection, we will have a mating pool of size N/2 – 0.1N to recombine to create a new generation. This process is similar to how a child will inherit certain biological features from his/her parents. We select 2 parents randomly and combine their DNA using the uniform crossover method. That is, for every bit of the child's DNA sequence, there exists a 50% chance that the feature was inherited from Parent 1, likewise for Parent 2. We allow each pair of parents to produce 2 children for further exploration of the optimal subset. Naturally, alteration in genes can happen spontaneously. Mimicking this, we also let each offspring undergo the mutation process. That is, every feature of the child has a 1/m chance of being changed from "0" to "1" or vice versa.

Finally, the above steps will form a loop. Each generation of size N will recombine and produce a new generation of size N/2. Our stopping criteria for this algorithm is when there is one remaining individual left. This last individual represents the optimal feature subset of our data that will form the ANN + GA model to be trained similarly to that described in Section 3.2.

3.4 Network Pruning

The pruning technique used in this research was that introduced by Gedeon, T. and Harris, D. (1992) in the paper "Progressive Image Compression". Aside from feature selection, the current baseline architecture is rather complex given the high number of hidden neurons in comparison to the number of input features. As there are 100 hidden units and 39 predictors, there are a total of 78 x 100 weights + 100 biases = 7900 free parameters for 368 training samples. Research done by Zhang et al. (2017) has shown that a two-layer ReLU model with 2n+d free parameters can, theoretically, fit any dataset of n patterns and d dimensions. Hence, it is more than possible that our current baseline/subset models can lead to overfitting and poor generalisation.

Our approach to overcome this problem is to remove certain hidden neurons based on their distinctiveness which can be thought of as the unique contribution of each hidden unit to the final prediction of the model. This individual functionality, as described by Gedeon and Harris, can be computed by using the activation vectors of all hidden neurons over the pattern presentation set. Following this, we calculate the angular separation between every pair of activation vectors. This set of angles provided us with information on the contribution of each hidden neuron relative to other units. We then evaluate and prune the model based on the removal criteria as follows:

- Pairs with angular separation less than 15 degrees will be considered similar. One unit will then be removed and have its weights added to the other one.
- Pairs with angular separation more than 165 degrees will be considered complementary and both removed.

We start off with the minimum angle between activation vectors to be 15 degrees and the maximum to be 165. Once no additional unit is found, the lower threshold will then be increased by 10 degrees incrementally (if needed) until the desired sparsity is reached (95% was the final sparsity used in this paper). The pruned network will then be retrained and tested with unseen data whose accuracy will be used to evaluate the performance of the network before and after pruning. Once again, the validation method used is leave-one-out-participant. The result reported is an averaged value across all validation sets.

3.5 Hyperparameter Tuning

The hyperparameters associated with three processes as described above were incrementally tested. The validation method used for hyperparameter tuning is 80/20.

A. Classification by baseline ANN:
- Adam Optimiser's learning rate
- No of epoch

B. Feature Selection By GA:
- Selection Pressure

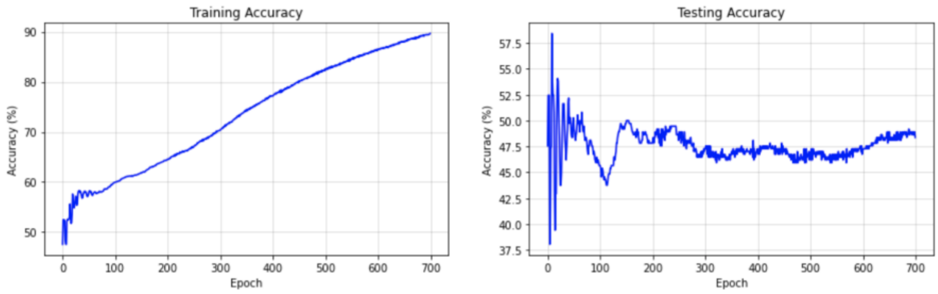
C. Post-pruning classification:
- Adam Optimiser's learning rate
- No of epoch

4 Results and Discussion

4.1 Classification result of the the baseline ANN

Using the baseline ANN, the result averaged over all validation set is 48.4%, lower than both the result achieved by Zhu and the chance level of 50%. Notably, the results vary substantially between validation sets as the best accuracy achieved is 68.75% while the worst is 6.25%. This raises the question of whether there exist some inconsistencies between observers being better at distinguishing untruthfulness than others. Furthermore, as we plot the averaged training and testing accuracies over all epoch, it can be seen that the testing accuracy almost reached the 50% level at around 200-250 epoch and subsequently levelled out to approximately 47.5% when reaching 300 epochs, showing signs of overfitting. This behaviour is expected due to the complex baseline architecture comparing to the modest number of training samples.

Fig. 3. Training and testing accuracies for baseline ANN model (averaged over all validation sets)



However, a closer inspection shows that the ANN does perform better in identifying contents without manipulated beliefs. That is, the neural network correctly predicts when the speaker did believe in the materials they presented with an accuracy of 57%, higher than that when there is manipulated beliefs, 50.5%). This is consistent with the result obtained in Zhu's paper. Such findings are rather interesting as it suggests that while our pupillary responses cannot directly help us in identifying lies, they can help decide who/what to trust. More research needs to be done, however, to make any concrete statement regarding this conclusion as pupillary signals can be easily affected by environmental factors (e.g. room temperature, tiredness/decreased attention in the observer after watching many videos continuously). Indeed, research has shown that an increased cognitive load can lead to pupil dilation (Steinhauer et al., 2004).

4.2 Results after Feature Selection

Using Genetic Algorithm for feature reduction, the subset of features attained include the followings:

1. variance pupillary size (PS) of the left eye
2. root mean square PS of the left eye
3. means of absolute values of first differences PS of the left eye
4. mean pupillary size for right eye
5. means of absolute values of first differences PS of the right eye
6. means of absolute values of second differences PS of the right eye
7. amplitude of the peak occurrences in the VLP PS signals of the right eye
8. amplitude of the peak occurrences in the LP PS signals of the right eye
9. ratio of peak occurrences between VLP to LP signals for right eye
10. minimum averaged PS
11. standard deviation of the averaged PS
12. variance of the averaged PS
13. root mean square of the averaged PS
14. means of absolute values of first differences PS of the averaged PS
15. means of absolute values of second differences PS of the averaged PS
16. number of peak occurrences in the VLP averaged PS signals

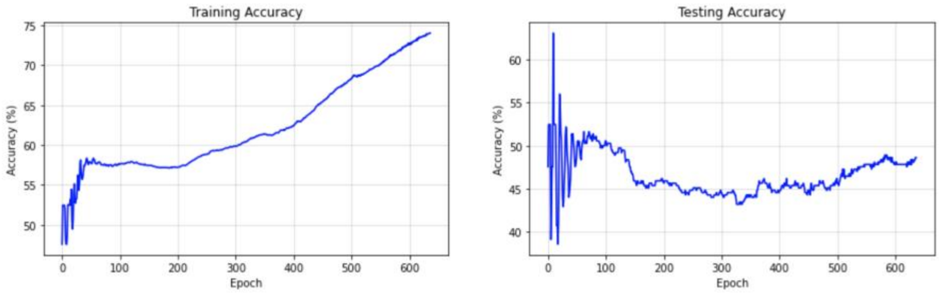
It appears that one of the more important features involve the ranges or how spread out was the pupillary size across the viewing of one video. Furthermore, the gradual changes in pupillary sizes appear to have a large influence on the neural network ability to detect deception as the means of the first and second differences for both left and right PS were retained by our method. Information about the peak occurrences is also deemed relevant for our model by GA.

We see only a slight improvement as the testing accuracy averaged over all validation sets is now 48.6%. This result is lower than expected as we anticipated a reduced number of input features would substantially increase the model performance. Hence, we suspect the problem of overfitting is still present. Once again, the GA + ANN model predicts truthfulness in recordings significantly better than otherwise. Detailed accuracy can be seen in Table 1.

Table 1. Results using baseline ANN and ANN + GA compared to that obtained by Zhu.

	Correct Doubt	Correct Trust	Total
Baseline ANN	50.5%	57%	48.4%
Baseline ANN + GA	49.3%	52.5%	48.6%
Zhu's result	50%	66.7%	58.3%

Fig. 4. Training and testing accuracies for baseline ANN + GA model (averaged over all validation sets)

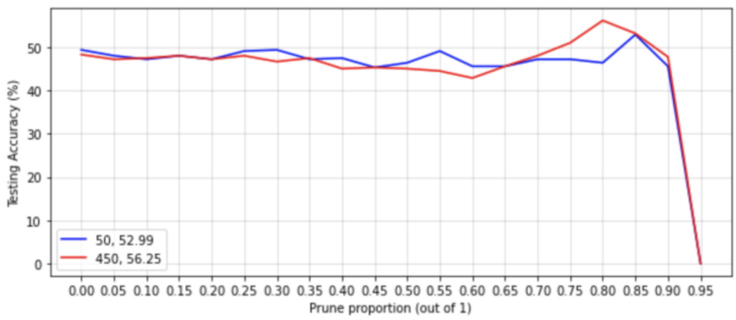


4.3 Results of Network Pruning

In this section, we will look at the pruned model performance with re-training and compare the effect of pruning on both the baseline model and the model with subset features selected by GA in section 4.2.

For the ANN model, as we experiment with numerous combinations of Adam's Optimiser learning rate and the number of epoch for training, we observe that the testing accuracy stays relatively constant around the pre-pruning level of 48.4%. It then generally peaks up to the best result, which can be as high as 56.25%, when 80-85% of the hidden neurons were removed. After this, the model performance gradually decreases in performance and plunges to 0% as we eliminate 95% of the hidden units. This behaviour is expected due to the overfitting problem as we have discussed. By reducing the number of free parameters, we allow the model to improve its approximation on important and distinctive nodes using the limited information from the training set without worrying about other non-unique nodes. Hence, not only this simplified model of 10 hidden neurons can reduce training time, but it has proved to also enhance model generalisation.

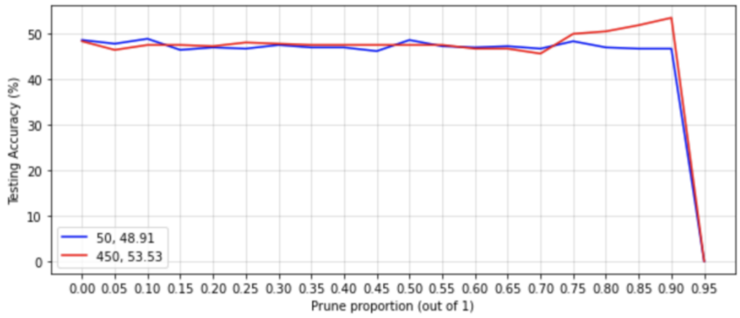
Fig. 5. Testing accuracies for the baseline ANN model over the proportion of hidden neurons pruned (averaged over all validation sets)



Note. The blue line represents for the pruning result of 50 retraining epoch, while the red line shows that of 450 retraining epoch. The legend includes the number of epoch used and the best accuracy achieved.

For the ANN + GA model, we see a similar trend where the testing accuracy remains fairly constant within the pre-pruning accuracy of 48.6%. It then once again peaks up at around 90% sparsity. This value can be as high as 53.53%. Similarly to the pruned ANN model, the simplified ANN + GA model only starts to lose performance from 90% sparsity and dove to 0% accuracy at 95% sparsity. Given that the number of input predictors is half of that in the baseline model, it is reasonable that more hidden units were pruned for the ANN + GA model.

Fig. 6. Testing accuracies for the ANN + GA model over the proportion of hidden neurons pruned (averaged over all validation sets)



From the above results, we have successfully found two simplified ANN model designs suggested both by GA and Network Pruning. The best models' performances achieved from these suggestions are detailed in Table 2 in the Appendix.

4.4 Limitations

The first apparent problem that we have encountered is the modest number of training samples comparing to its high-dimensional nature which has led to the models struggling to perform well on unseen data. Another important factor pointed out by Zhu was the possible inconsistencies regarding certain presenters being better deceivers than others. More importantly, while we have explored extensively the contributions of GA and hidden unit pruning in using one's pupillary size in distinguishing dishonesty, the achieved results stayed relatively around the chance level. This raises the question of whether or not pupillary responses alone can be indicative of one recognising untruthfulness. As mentioned in Section 4.1, one's pupillary features can be easily affected by room temperature, video content, level of attention, etc. Therefore, it can be helpful to expand our research to the rest of the available dataset (including the features of Blood Volume Pulse, Galvanic Skin Responses, and Skin Temperature) to fully capture the observers' physiological state, despite our original intention of focusing on pupil dilation.

5 Conclusion and Future Works

This paper has attempted to re-investigate the possibility of recognising dishonesty in a recording through an observer's pupillary responses, via the means of an ANN. Our baseline ANN has shown to predict untruthfulness accurately 48.4%, less than that of a flip of a coin. We raised the potential issue of overfitting due to the overly complicated baseline model design. As an attempt to simplify the structure of the neural network, we first used Genetic Algorithm to select the most relevant features of the dataset. It appears that the models' performance relied on three important information: 1) the range of pupillary size throughout the viewing of one video, the gradual change (reflective in first and second differences), and the peak occurrences. Using this ANN + GA model, we reduced the number of input features down to 12 and obtained a slightly better result of 48.6%. The second part of this attempt was hidden unit pruning. The decision regarding which unit to eliminate was based on the unique functionality of each neuron and its contribution to the model's final prediction. It was shown that by removing 80% of the hidden neurons for the baseline ANN and 90% for the ANN + GA model, we increased the model accuracies to 56.3% and 53.5%, respectively. It is notable, moreover, that the models consistently predicted truthfulness more accurately than dishonesty.

Numerous factors can be improved in future works. With regards to experimental design, it may be beneficial to spread out the observer experiment to multiple sessions and consult with an optometrist to better understand the properties of the eyes when losing concentration or detecting things/facial expressions of significant importance. It is also crucial to stay consistent with the presenters and their performance abilities. Furthermore, we can broaden the scope of our dataset to other types of physiological signals to more comprehensively capture one's mental processes during identifying dishonesty. As this suggestion once again introduces new predictors, more data points are needed for our training data to be representative of the population.

Concerning model design, we can experiment with model architecture by adding extra hidden layers or using different activation functions. Finally, regarding network reduction, we can explore hidden layer pruning based on other properties.

References

1. Gedeon, T., Harris, D. (1992). Progressive image compression. IJCNN International Joint Conference On Neural Networks. Retrieved from: https://doi.org/10.1109/ijcnn.1992.227311

2. Gomez, F., Quesada, A., Lopez, R. (2020) Genetic Algorithms for Feature Selection. Neural Designer. Retrieved from: https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection

3. Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. International Journal Of Human-Computer Studies, 59(1-2), 185-198. Retrieved from: https://doi.org/10.1016/s1071-5819(03)00017-x

4. Prochazkova, E., Prochazkova, L., Giffin, M., Scholte, H., De Dreu, C., & Kret, M. (2018). Pupil mimicry promotes trust through the theory-of-mind network. Proceedings Of The National Academy Of Sciences, 115(31), E7265-E7274. Retrieved from: https://doi.org/10.1073/pnas.1803916115

5. Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. International journal of psychophysiology : official journal of the International Organization of Psychophysiology, 52(1), 77–86. Retrieved from: https://doi.org/10.1016/j.ijpsycho.2003.12.005

6. Zhu, X., Gedeon, T., Caldwell, S., Jones, R., Gu, X. (2020). Deceit Detection: Identification of Presenter's Subjective Doubt Using Affective Observation Neural Network Analysis. 2020 IEEE International Conference On Systems, Man, And Cybernetics (SMC). Retrieved from: https://doi.org/10.1109/smc42975.2020.9283210

7. Zhu, X., Qin, Z., Gedeon, T., Jones, R., Hossain, M., Caldwell, S. (2018). Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers' Pupillary Responses. Neural Information Processing, 610-621. Retrieved from: https://doi.org/10.1007/978-3-030-04212-7-54

Appendix

Table 2. Hyperparameter settings.

	Baseline ANN	**ANN + GA**	**Baseline ANN + Network Pruning**	**ANN + GA + Network Pruning**
No. of input features	39	16	39	16
No. of hidden neurons	100	100	20	10
GA Selection Pressure		1.2		1.2
Initial training Adam's Optimiser Learning rate	0.0013	0.0013	0.0013	0.0013
Initial training No. of epoch	700	637	700	637
Post-pruning Adam's Optimiser Learning rate			0.005	0.005
Post-pruning No. of epoch			450	450
Accuracy	48.4%	48.6%	56.3%	53.5%