Evolutionary Algorithm with Functional Measurement as pre-processing: A Hybrid Feature Selection Model

Jiyang Zheng[1]

The Australian National University, Canberra ACT 2601, Australia
{jiyang.zheng}@anu.edu.au

Abstract. Feature selection, as a widely-used data pre-processing technique, has played an important role in traditional machine learning tasks on improving the model performance and reducing the computational complexity. This idea has also been extended to the deep learning and neural network fields, where evidence shows that feature selection could generalise the network via reducing the influence of input to the output. In this paper, we introduce a hybrid model for feature selection on a back-propagation trained neural network where the model uses the functional measure to reduce the dimensionality in the original dataset and then an evolutionary algorithm to select the best subset of features for a given classification problem on facial features. The neural network that uses the hybrid model has achieved an testing accuracy of 81% and an F1 score of 0.88 using k-fold cross-validation. It has also outperformed the baseline models by decreasing the accuracy gap between training and testing set which reduces the model overfitting problem.

Keywords: Feature Selection · Neural Network · Evolutionary Algorithm

1 Introduction

When training the neural network over large datasets with high dimensionality, it is quite often that the neural network gets overfitting to the train samples which cause the model to have lower generalisation and performance. To solve this problem, a naive approach will be to let the neural network learn less about the train samples and ideally to only capture the most important aspects of the data. In this paper, we have introduced a hybrid feature selection model. The model first pre-processes the dataset by applying a filter method to the input features which uses the functional measure to remove a subset of features. After that, the model uses an evolutionary algorithm to select feature subsets from the reduced dataset and evaluate the performance of feature subsets by measuring the testing accuracy and F1 score of the neural network trained by the subset. The function measurement uses the distinctiveness analysis [5] as the proxy to measure the similarity using angles between multi-dimensional features. It is an extension to the Garson saliency measures [6], which enable the functionality of hidden neurons to be measured via the pattern matrix [4]. In this paper, we will refer to this functional measure as *Model I.*

In this paper, we have compared three models which are the conventional Neural Network, the Neural Network with functional measure only and the Neural Network with functional measure and evolutionary algorithm (The hybrid model). The task we choose to evaluate the model is a binary classification task with a set of facial features. The network output should be either 1 represents the given sample contains facial features of the same person or 0 represents the given sample contains facial features of two different peoples. The data set itself has 12 sets of 3 images, where two of them in the same set are determined to be the same person, and the other is identified as another person. The facial features data contains two sets of numerical data which provides the x,y coordinates of a facial feature in pairs of the two photos out of three [1] and the distance between two facial features from the same photo[2], also in pairs [2].

1.1 Data Inspection

The selected data contains two different datasets which measure the historical facial images from different perspectives. We first investigate the ground-truth value distribution of the dataset (Figure 1). Wherefrom the diagram, it is observed that the two datasets have the same class distribution. The sample points belong to class zero is has doubled the amount of the sample points of class one. Therefore, since the dataset is unbalanced, we would expect a certain level of bias of the trained model. There might be a situation where the accuracy score is high due to the model naturally predict most sample to class 0. To avoid the mismeasurement of the model performance. In addition to the accuracy scores, we will also investigate the **F1 score** of the results to cross-evaluate the model.

[1] FFMs dataset
[2] Distance dataset

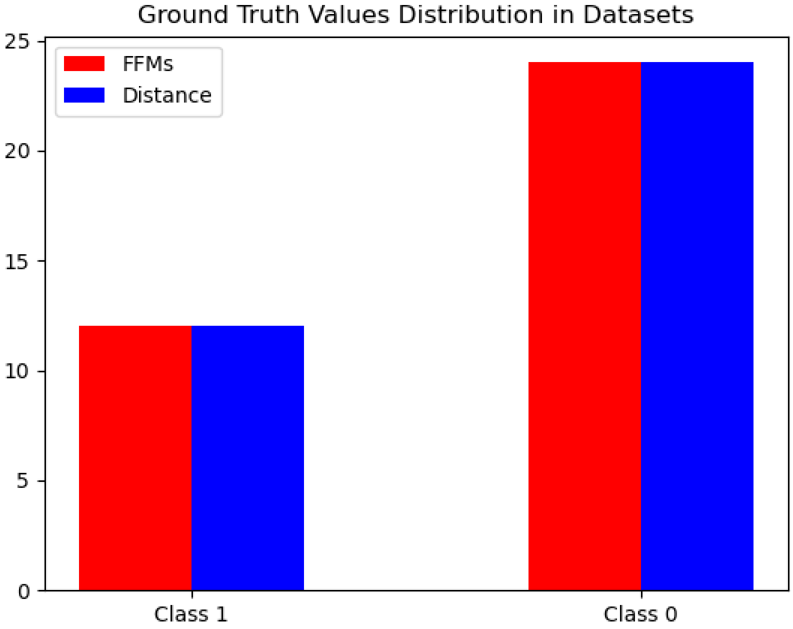


Fig. 1: Class Distribution in FFMs and Distance

From the description of data, the two datasets contain 36 data samples each and with 57 and 183 feature respectively. We visualise the distribution of the average values of each feature in the dataset (Figure 2). The FFMs value represents the coordinate position with respects to the x and y-axis, the median is around 180 and the overall distribution is right-skewed. The average features value of distance has shown a distribution that is close to a normal distribution where the median is in the middle of the box and both tails are at a similar length. Therefore, from a statistical view, the distance dataset might be better in representing the historical image as the average feature values distribution is more evenly spread.

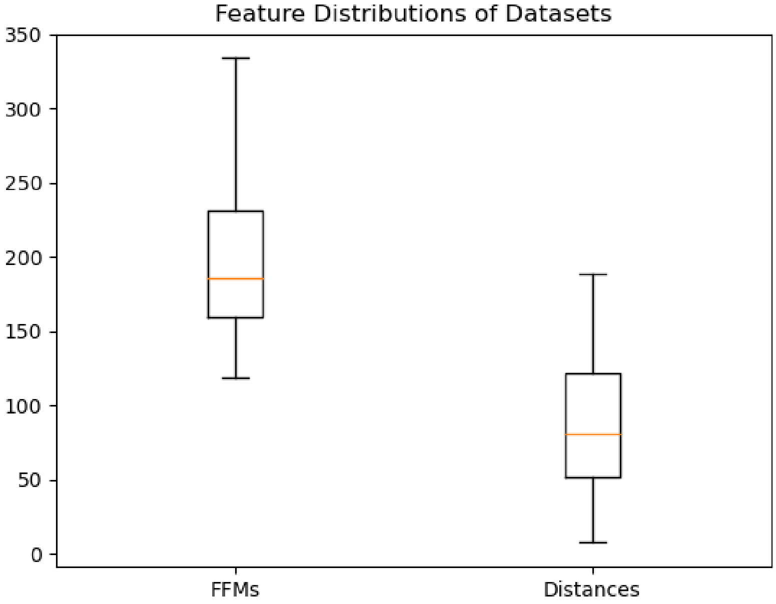


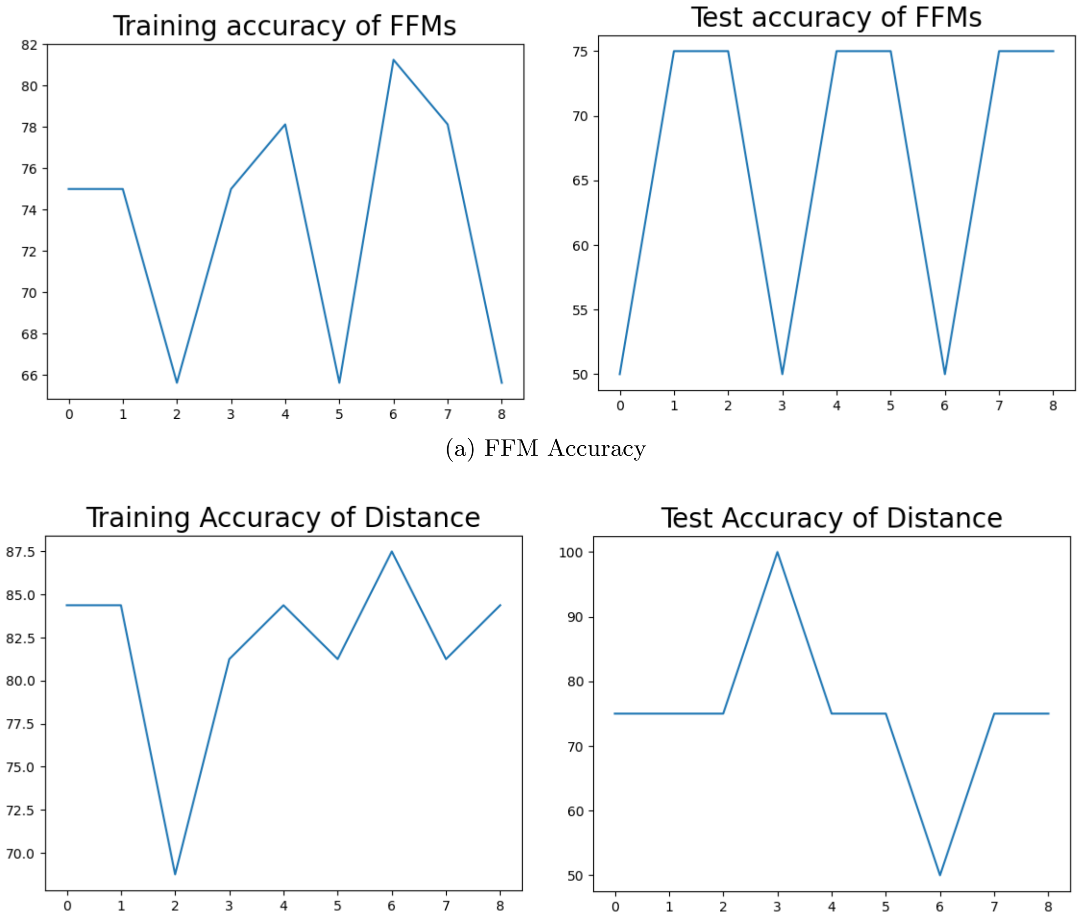
Fig. 2: Average Feature Value Distribution of FFMs and Distance

1.2 Evaluation method

The evaluation method uses k-fold cross-validation to measure the performance of the feature selection models. The configure of k is determined by the number of sample data in each dataset. From the description of the datasets, we know that there are 36 distinct data samples in each dataset and the number of features is 57 and 183 respectively. Since the dimensionality is relatively large with regards to the number of sample data we have, we would intend to keep more data for training the model, while the biased class distribution problem could be generally minimised by using cross-validation. Therefore, we select the k as 9 which splits the data into a training set with 32 data samples and 4 data samples for the testing set, where we have kept the most data sample as the training set with a proportion of 8 : 1. In the final confusion matrix for calculating the F1 score, we will expect a total number of 288 training predictions and 36 testing predictions.

1.3 Data Selection using Conventional Neural Network

To maintain the consistency of our measurement and makes the evaluation simpler to understand. We will select one of the two datasets for our later model training and testing. From our above observation, the ”Distance” dataset would better represent the image. To validate our hypothesis, we constructed a conventional neural network using both of the datasets, with the evaluation method introduced above, we measure the accuracies of the model in each validation trail. From the (Figure 3), it is observed that the model trained by the ”Distance” dataset has a more stable performance as the fluctuation of accuracies are smaller. The statistical summary of the performance is shown in (Table 1).



(a) FFM Accuracy

(b) Distance Accuracy

Fig. 3: Training and Testing Accuracies

Table 1: Statistical Summary

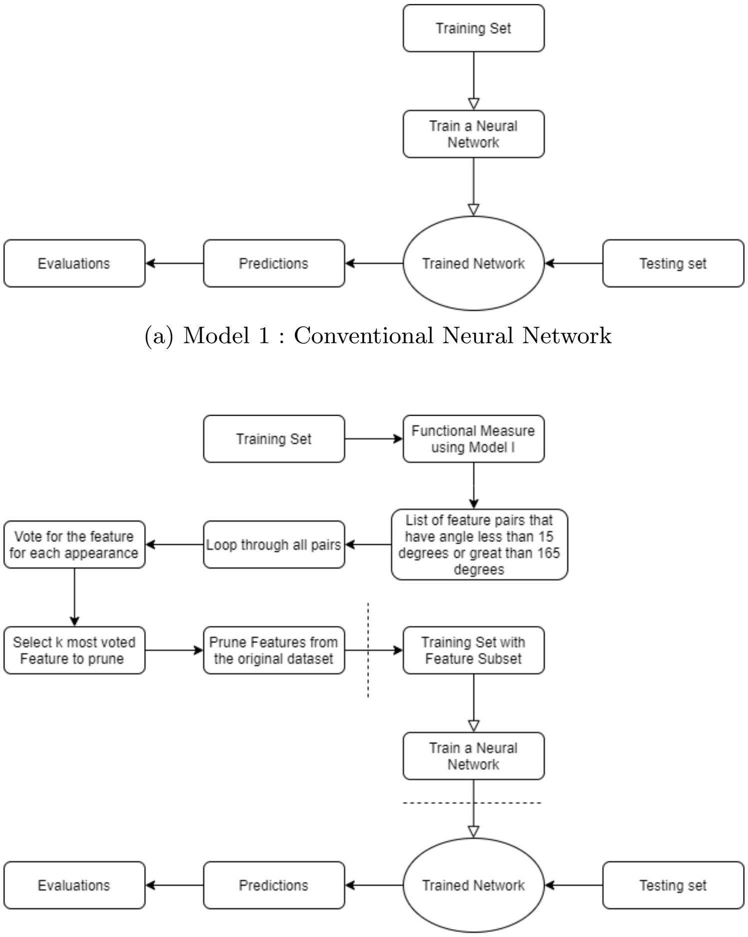
Classification Performance Measure	FFMs Training	FFMs Testing	Distance Training	Distance Testing
Average Accuracy	73.3	66.7	85.07	75
F-score	0.83	0.8	0.9	0.84

The general model performance with the Distance set has outperformed the model with FFMs set, which is consistent with our hypothesis from the feature distribution analysis. However, it is noticeable that both models are largely overfitting, therefore, we need to apply feature selection to prevent this problem. In summary, we will use the "Distance" dataset to train the backpropagating neural network with different models which will be introduced later in the paper.

2 Methods

2.1 Model Pipelines

First, we will introduce the general pipeline of the three models with a general pipeline of how the model is constructed. Figure 4 shows the two baseline models which are: A conventional neural network, where the data is directly used to train the neural network without any pre-processing step; and a neural network with functional measure, where the data is pre-processed by distinctiveness measure using Model I introduced in [1] and use the pruned training set to train the neural network for classifications.


Training Set
Train a Neural Network
Trained Network
Testing set
Predictions
Evaluations
Training Set
Functional Measure using Model I
List of feature pairs that have angle less than 15 degrees or great than 165 degrees
Loop through all pairs
Vote for the feature for each appearance
Select k most voted Feature to prune
Prune Features from the original dataset
Training Set with Feature Subset
Train a Neural Network
Trained Network
Testing set
Predictions
Evaluations

(a) Model 1 : Conventional Neural Network

(b) Model 2 : Neural Network with Functional Measure

Fig. 4: Baseline Models

The pipeline of the hybrid model is shown in Figure 5. The model has used functional measure as a pre-processing technique, the dataset with reduced dimensionality will be sent to the evolutionary algorithm, where the algorithm will do the feature selection again on the dataset and select the optimal feature subset to be the final feature set. We will use the final feature subset to train our neural network and evaluates its classification performance

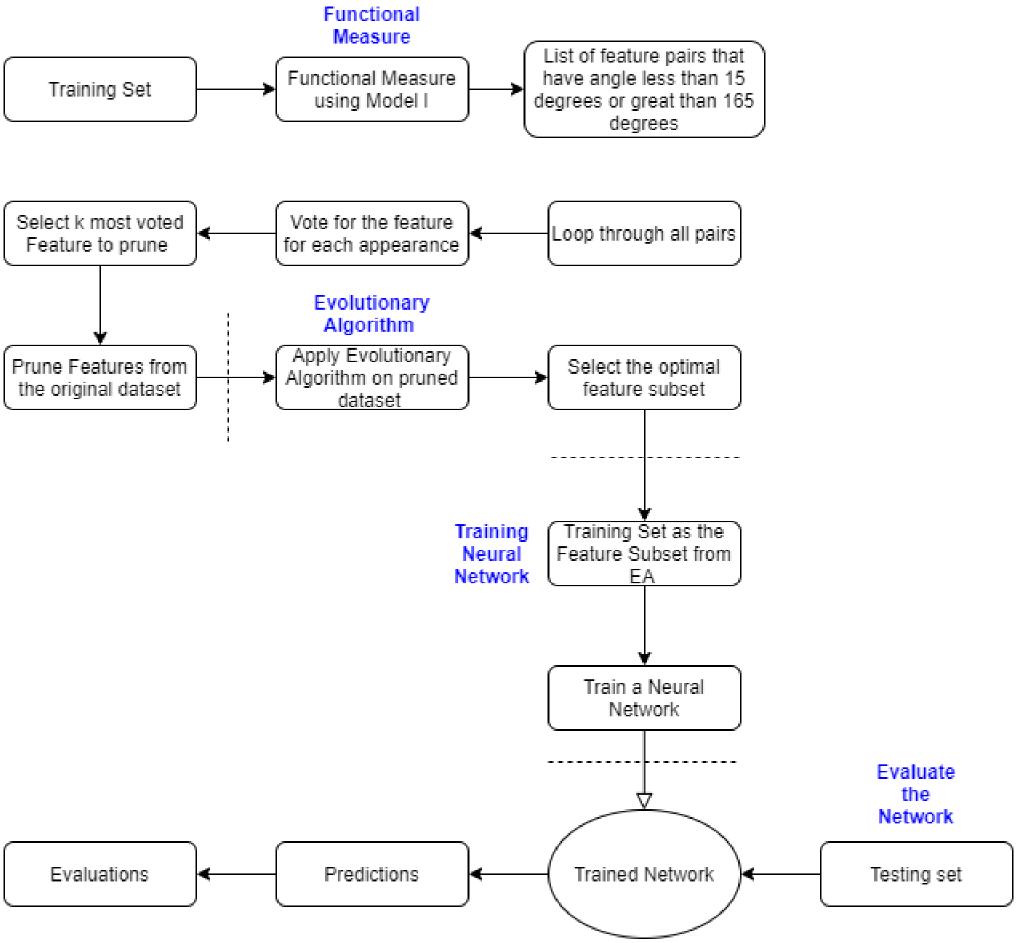

Functional Measure
Training Set
Functional Measure using Model I
List of feature pairs that have angle less than 15 degrees or great than 165 degrees
Select k most voted Feature to prune
Vote for the feature for each appearance
Loop through all pairs
Evolutionary Algorithm
Prune Features from the original dataset
Apply Evolutionary Algorithm on pruned dataset
Select the optimal feature subset
Training Neural Network
Training Set as the Feature Subset from EA
Train a Neural Network
Evaluate the Network
Evaluations
Predictions
Trained Network
Testing set

Fig. 5: The Hybrid Model

2.2 Implementation Details

Model Selection The model selection refers to the machine learning model we use to evaluate our feature selection model, for this paper, we have tried two different models on predicting the facial feature image with the Distance dataset, which is the conventional neural network classifier and the logistic classifier. The evaluation is shown in table 2. The results show the good accuracy on the training performance using the Logistic model, however, the testing performance is far worse than the training performance, which shown the model is largely overfitting and the overall performance of Neural Network is relatively stable. We will use the Neural Network as our machine learning and discuss the potential of Logistic Classifier later in the future work part.

Table 2: Neural networks Classifier and Logistic Classifier

Classification Performance Measure	Logistic Training	Logistic Testing	NN Training	NN Testing
Average Accuracy	1	61	85.07	75
F-score	1	0.66	0.9	0.84

Hyper-parameters The hyper-parameter for the conventional neural network include the number of hidden layers, the number of hidden neurons in each layer, the learning rate and so on. For the functional measure, we need to decide on the number of features we want to prune and the model of distinctive measure. And the evolutionary algorithm has the mutation probability as well as convergence trail for the hyper-parameter. The below Table 3 shows our selection of the hyper-parameters.

Table 3: Hyper-Parameter Table

Hyper-parameter	Model	Selection
Hidden layer	Neural Network	1
Hidden Neurons - Layer 1	Neural Network	30
Learning Rate	Neural Network	0.01
Optimiser	Neural Network	SGD
Loss Function	Neural Network	Cross-Entropy Loss
Number of prune Feature	Functional Measure	36
Distinctive Measure Model	Functional Measure	Model I
Crossover Probability	EA	0.8
Mutation Probability	EA	0.2
Population	Evolutionary Algorithm	30
N-generation	Evolutionary Algorithm	400

Functional Measurement We first train the back-propagate neural network without any feature selection technique applied to get the converged weights and activation neurons. In this paper, we used two different functional measurement known as the I and W mode [1] where they find the angle between each pair of pattern or weight to determine whether two input vectors are considered as high correlated input. The angle measurement formula is given by [1]:

$$angle(m,n) = \cos^{-1}\left(\sqrt{\frac{\sum_p \mathrm{sact}(p,m)^2 \cdot \sum_p \mathrm{sact}(p,n)^2}{(\sum_p \mathrm{sact}(p,m) \cdot \sum_p \mathrm{sact}(p,n))^2)}} - 1\right)$$

where
$\mathrm{sact}(p,i) = \mathrm{normalise}(\mathrm{weight}(i)) - 0.5$ for model W and
$\mathrm{sact}(p,i) = \mathrm{normalise}(\mathrm{pattern}(i)) - 0.5$ for model I

Via the measurement on the selections, it is observed that the angle between weights are all within the range of 15 and 165 degrees, where the range is $[52.7, 131.9]$, therefore, the distinctiveness measured by Model W isn't useful to this dataset compared to the other models. On the other hand, Model I has measured the angle between pattern with the range of $[2.3, 74.7]$ where we find there are many patterns pair that have an angle less than 15 which are considered as very similar. Hence, for the distance dataset, we use Model I as the functional measure model.

Pruned Neural Network By feeding the pattern to Model I, we can get a list of vector pairs that have angles less than 15 degree, which means the information they contain is similar. Since one will capture the most information of another, pruning one of the pairs will not have a significant loss of information. In the algorithm, we used the technique of voting, where we count the number of the appearance of the feature in the pairs selected by Model I. For example, we extract a pair of vectors that has an angle less than 15 degree and add one to its count, we repeat that until all the pairs are accessed. After that, the k features with the most voting will be identified as the least informative features. After getting the index of the least significant features, we remove those features from the original dataset and use the new reduced data set to re-train the neural network from scratch. Finally, we access the performance of these baseline neural networks by using the testing set of data. The reduced dataset is also passed to the evolutionary algorithm for constructing the hybrid model.

Evolutionary Algorithm The evolutionary algorithm takes the dataset that has been pruned by the functional measure and does the feature selection again on the reduced dataset [8]. A general procedure of the algorithm follows as below:

1. Use Binary chromosome to represent a feature subset.
2. Extract the actual feature subset with data samples via finding the index of ones in the binary chromosome representation and refer them to the Distance dataset
3. Initialise the population, we have chosen $n = 30$ due to the time limitation.
4. Evaluate the feature subset by training a neural network classifier with the actual feature subset extracted in step 2 and evaluate its testing accuracy.
5. Apply single-point crossover to the chromosome representation
6. Apply Multiple flip bit mutation on the chromosome representation with a probability of 20%
7. Extract the actual feature subset with data samples with evolved chromosome representation
8. Build and evaluated the model with the new extracted data subset
9. Select the chromosome that will proceed to the next generation using tournament selection. The higher the accuracy, the better the fitness value.
10. Repeat Step 5-9 until the number of generation has been reached

Due to the limitation of computational power, we have restrict the population and number of generation to a small number where it could be adjust to a larger number for better convergence. The flip-bit mutation has simulate the process of adding or deleting features from the feature subset and the crossover can be thought as the two feature subsets swapping their features.

3 Experimental Results and Discussion

3.1 Experiments and Results

We have trained three models with the Distance dataset and evaluate them using 9-fold cross-validation. The conventional neural network appears to have a relatively stable performance with low accuracy and F-score on its testing set, which means the model is largely overfitting to the training sample (Figure 6).

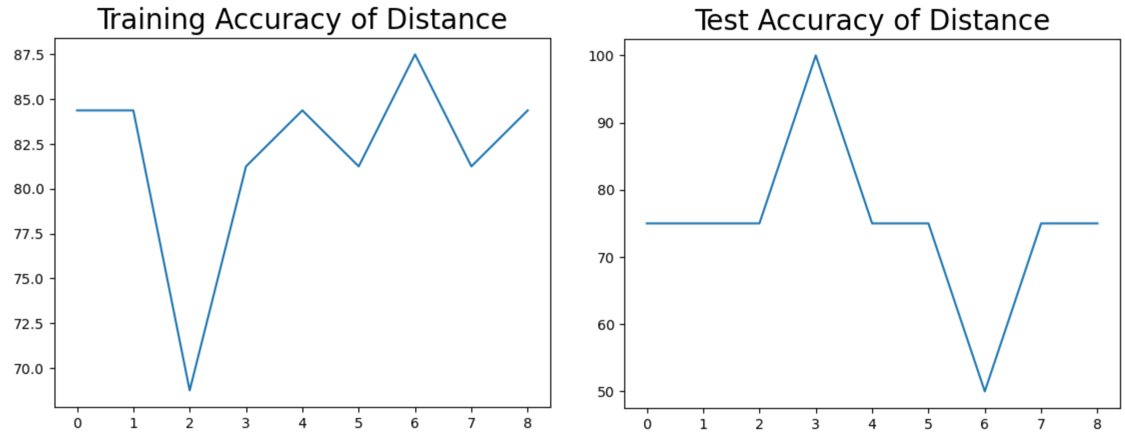


Fig. 6: Training and Testing Accuracies of Conventional NN in 9-fold Cross Validation

The training loss is shown in (Figure 7), the lowest loss is around 0.38 and as we can see there are a lot of fluctuations in the loss trend lines, which means the dataset are strongly biased which has caused the optimiser to jump out of the local minimum. The neural network trained with functional measure and pruned dataset has its loss trend lines closer between each validation runs. Among the three models, the neural network with the functional measure has appeared to be the most stable model with a relative high accuracy score. The training accuracy has dropped by 2.43%, however, the testing accuracy has been raised by 2.78%. The trend of its accuracy is shown in the Figure 8. Due to the expensive time consumption for each run of evolutionary algorithm, we have only tested the hybrid for a handful of times, the performance of the hybrid model isn't as stable as the two models mentioned above, mainly due to the randomness of the algorithm and we have set the population number and number of generation to a small number. Ideally, if we increase number of repetition to convergence, the performance will become more stable. However, using the hybrid model, we have achieved the most numbers of 100% accuracies. (Figure 9)

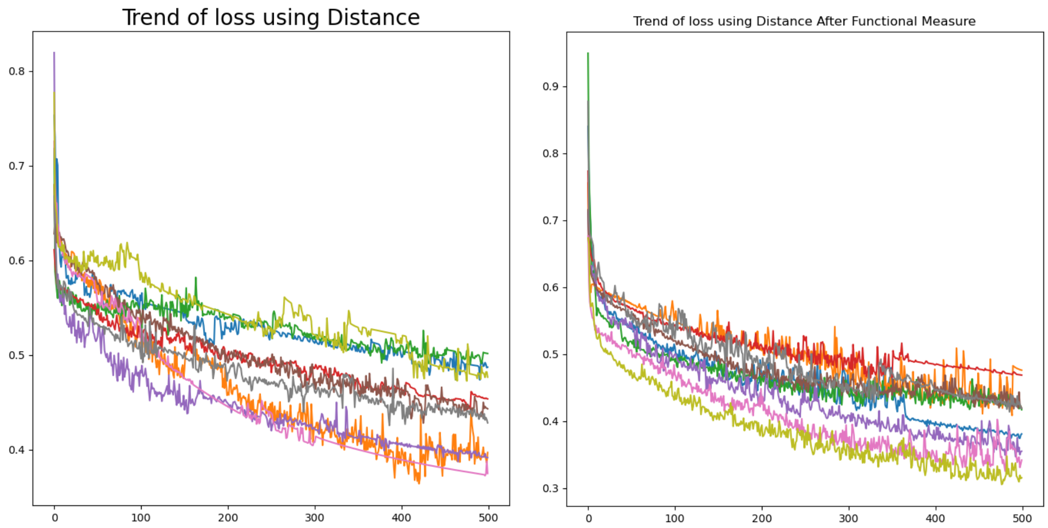


Fig. 7: Training Loss of Conventional Neural Network (Without & With Functional Measure)

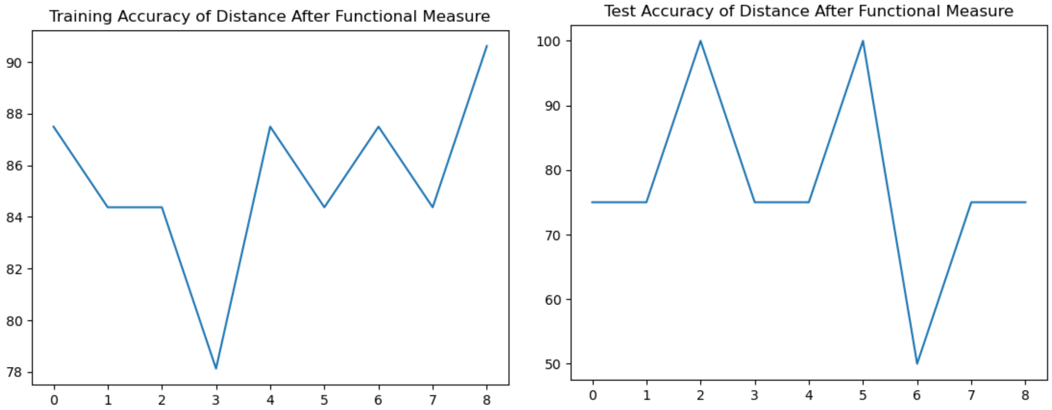


Fig. 8: Training and Testing Accuracies of NN with Functional measure in 9-fold Cross Validation

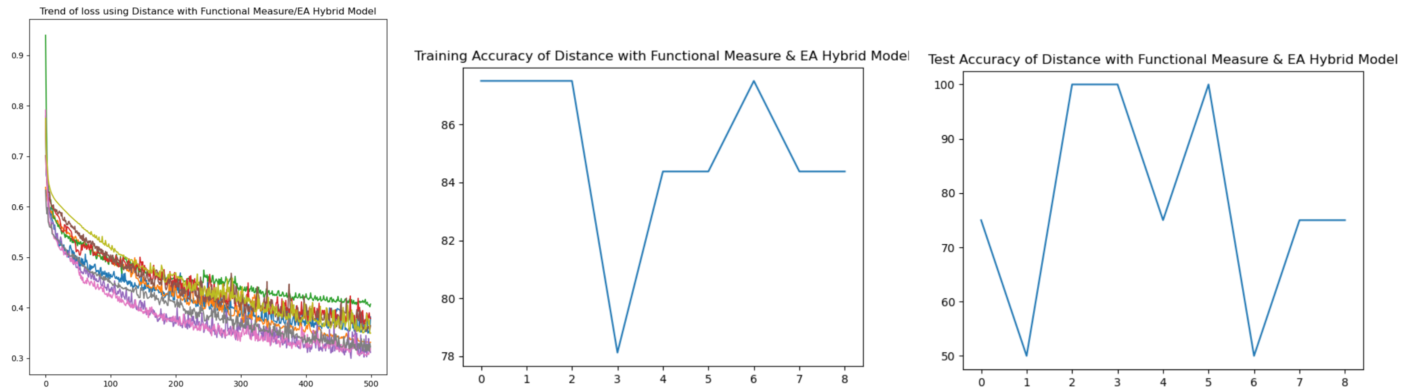


Fig. 9: Training Loss and accuracy Trend; Testing Accuracy Trend of the Hybrid Model

The overall performance of the three model are shown in the Table 4

Table 4: Statistical Summary

Classification Performance	NN Train	NN Test	FM Train	FM Testing	Hybrid Train	Hybrid Test
Average Accuracy	85.07	75	82.64	77.78	83.33	80.56
F-score	0.9	0.84	0.88	0.85	0.89	0.87

3.2 Discussion

From the experimental result, it is observed that the pruning of data has slightly improved the test accuracy for classification and reduce the gap between the training accuracy and testing accuracy which prevents the overfitting problem, where this as expected because the dataset contains a large number of features where the sample size is relatively small compared to the number of features, this would simply cause the model to over-training by the features. A dimensionality reduction would largely prevent the model overfitting and give a better generalisation to the model. The number of pruned features would also affect the model performance, it is found that a larger number of pruned features would increase the efficiency of the hybrid model which would allow a longer searching period therefore a chance of improving the general accuracy. However, it is observed that when the pruned features are more than 20% of the original features, the dataset will start to lose important information, this is shown by the major reduction in the training accuracies when the hyper-parameter for prune neuron is set to greater than 36. By the analysis to the confusion matrix of the classification result, it is observed that there are always more False-negative than False positive, this is caused by the bias of the dataset which is also mentioned in the data inspection section, there are over $\frac{2}{3}$ of data samples belong to the class 0, therefore, the model might be trained with bias too. The last problem is that it is noticed that the range of the testing accuracy is $[50, 100]$ for the hybrid model which shows the lack of stability of the model. This is mainly due to the nature of the evolutionary algorithm, where the optimal solution is not guaranteed, in addition to that, the algorithm will require a large number of iterations to converge to a sub-optimal solution.

4 Conclusion and Future Work

In this paper, we have evaluated two different feature selection models with their neural network classifiers. The performance of the neural network with a hybrid feature selection model is contrasted with a conventional baseline neural network and the neural network with functional measure only to discover the effect of feature selection using an evolutionary algorithm. It is discovered that the evolutionary algorithm could further reduce the model overfitting problem and bring better generalisation to the neural network. To further validate the result, we tend to define the evolutionary algorithm to minimise the fitness value(i.e the accuracy). As a result, the technique decreases the model training and testing accuracy without generalising the model. For the future development of this hybrid model, there are three general aspects: Model Stability, Model Generalisation and Model interpretability. Model stability is to improve the stability of the hybrid model, a possible approach might via pre-determine the number of selected features in the feature subset and run multiple trials with a final aggregation. Model generalisation refers to test the model using different machine learning models, as mentioned in the model selection section, the logistic classifier has a larger gap between its training and testing accuracy where a 100% training accuracy might indicate the dataset is linearly separable, while it is worth a trial to see if the hybrid model could raise the testing accuracy to 100% as well. Model interpretability is the last important aspect for the research, as the information captured by the facial feature data are mostly representative and have their physical meaning, in other words, they are distinctive features at the level of human interpretation. It is quite important and essential to maintaining a level of interpretability [7] of the feature selection to let the people understand why the certain feature should be pruned from a human-understandable perspective.

References

1. Gedeon, T.D., 1997. Data mining of inputs: analysing magnitude and functional measures. International Journal of Neural Systems, 8(02), pp.209-218.
2. Caldwell, S. (2021) ”Human interpretability of AI-mediated comparisons of sparse natural person photos,” CSTR-2021-1, School of Computing Technical Report, Australian National University.
3. Zhang, G.P., 2000. Neural networks for classification: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(4), pp.451-462.
4. Gedeon, T.D., 1995, November. Indicators of hidden neuron functionality: the weight matrix versus neuron behaviour. In Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems (pp. 26-29). IEEE.
5. Gedeon, TD and Harris, D “Network Reduction Techniques,” Proceedings International Conference on Neural Networks Methodologies and Applications, AMSE, vol. 1, pp. 119-126, San Diego, 1991.
6. G. D. Garson, “Interpreting neural network connection weights,” AI Expert, pp. 47–51, 1991.
7. Sha, Y. and Wang, M.D., 2017, August. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 233-240).
8. Xue, B., Zhang, M., Browne, W.N. and Yao, X., 2015. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4), pp.606-626.
9. 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.
10. Abd-Alsabour, N., 2014, October. A review on evolutionary feature selection. In 2014 European Modelling Symposium (pp. 20-26). IEEE.
11. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H., 2017. Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), pp.1-45.
12. Kim, Y.S., Street, W.N. and Menczer, F., 2003. Feature selection in data mining. In Data mining: opportunities and challenges (pp. 80-105). IGI Global.