The Influence of Finite Datasets in Detecting Real Smiles on Traditional Neural Networks and LSTM Networks

Cindy Sue

Research School of Computer Science, Australian National University (ANU)
u5183013@anu.edu.au
May, 2020

Abstract. Artificial Neural Networks are a widely used algorithm and sit behind some of the most successful applications in artificial intelligence with the ability to recognise hidden patterns in data to solving complex problems. Motivated by this field in computer science, this paper looks at the performance of a traditional multilayer perceptron neural network model and a Long Short Term Memory deep neural network model in detecting smiles on pupillary response data, including a distinctiveness pruning technique to reduce model complexity. We found that no matter how advanced neural networks are in solving complex problems on sequential data, they lose performance on finite datasets. In all cases, our models did not achieve an accuracy above 50%. We also found that distinctiveness pruning can give us a simpler model at the expense of model prediction confidence. Finally, we conclude that pupillary response and gender information do not provide as much predictive power as we were hoping for.

Keywords: Classification • Distinctiveness Pruning • Genuine Smiles • Sequential Data • Long Short Term Memory • Recurrent Neural Network • Multilayer Perceptron • Finite Datasets

1 Introduction

In the 1970s, Paul Ekman identified six core emotions in human beings which since then, have led to scientists disputing the exact number of emotions which humans can express. In particular, emotion recognition has become a well-researched area due to the large number of potential applications in the real world such as to improve user experience, identify suspicious behaviour, or applications in education and gaming [1]. On the contrary, more research could be done in distinguishing between real and genuine or fake smiles.

The human smile is complex. We smile when we are happy, but we also unconsciously smile when we see others smile, are surprised, embarrassed, socially anxious and more [2]. The existence of this complexity in our emotions means that even the human vision system struggles to accurately distinguish between a genuine or fake smile. The ability to read the difference between a fake and a genuine smile can help humans to react to the situation and make decisions accordingly. In 2015, British psychologist Richard Wiseman said that the general public can differentiate smiles at around 60% accuracy and an individual's ability to tell the difference is also related to their profession and emotional intelligence.

Machine learning approaches have produced very promising results in distinguishing between genuine and fake smiles compared to humans. In 2007, Valstar et. al. [3] tried to differentiate real smiles from fake smiles through video using kernel methods and ensemble learning techniques which achieved 94% accuracy. A significant finding in their paper was that the head was the most reliable predictor, followed closely by the face. In 2019, Moussa et. al. [4] developed an artificial neural network model (henceforth referred to as a neural network [5]) to distinguish between real and fake smiles by using electroencephalograms (EEG) signals, achieving an accuracy of 78.29%. More recently, Hossain et. al. [6] found a strong association between pupillary responses in humans when presented with real smile and fake smile stimuli. They conclude that pupillary responses reflect a displayer's actual state of mind and judgement and gender is also significant to the pupillary differences observed. However, in their paper, they did not implement a machine learning approach.

Motivated by the success of neural networks in deception recognition [7], in this paper, we explore how well a simple multi-layer perceptron (MLP) neural network classification model can use pupillary response data to differentiate between genuine and fake smiles. We will refer to this MLP model as the baseline model. The data collected by Hossain et. al. [6] in their study of how pupillary responses in participants can help in smile detection will provide us with a good foundation for this paper. We will use the original pupillary

response dataset collected in their study, including inheriting the same definition of real and fake smiles from the paper. That is, a real smile is defined as a smile stemmed from happiness, and a fake smile is defined as one that has been posed, acted or one not a result of happiness. We then explore how much we can decrease the complexity of the MLP model by before sacrificing its performance using a distinctive network pruning technique on activation vector angles [8]. The pruning of neural networks generally help to decrease model complexity making it more interpretable and hence more desirable, increase generalisation to new data and increase inference speed. Finally, we explore what exploiting the sequential structure of the data in machine learning using a Long Short Term Memory (LSTM) [9] deep neural network classification model can have on model performance.

The contributions of this paper can be summarised as follows:

- We show that finite datasets heavily impact the performance of a neural network. Even though this is the case, the simple MLP and LSTM models are still able to provide predictions that are as good as random guessing. As a result, we did not see advantages in using a more complex LSTM model to capture memory of earlier sequential inputs.
- We show that pupillary response and gender information do not provide as much predictive power as we were hoping for compared to the work by Moussa et. al. [4] using EEG signals.
- We show that distinctive network pruning can give us a simpler model, but the trade-off here is the prediction confidence.

2 Methodology

In this section, we describe the attributes of the pupillary response dataset to be used for modelling as well as how it will be prepared to be used for building the MLP and LSTM neural network classification models since the input structure to these two models are different. The MLP model requires a fixed input structure compared to the LSTM model which accommodates input structures of varying length such as sequential data. We will then define how the MLP and LSTM models will be trained. In addition, this section will cover the distinctiveness pruning technique to be used on the trained MLP.

2.1 Dataset Description

The complete dataset captures pupillary responses from 10 healthy participants (observers) of Asian descent of which 6 were male and 4 were female [6]. In a controlled environment, the participants were presented with 19 grayscale video stimuli of fake and real smiles, each lasting 10 seconds long and adjusted to similar luminance ranges. Pupillary responses were collected throughout the whole video presentation with a cubic spline interpolation technique used to recover pupil size where the pupil was obscured due to blinking. Two versions of this dataset will be used in this paper – the raw dataset for the LSTM model, and an aggregated version for the MLP model.

The aggregated version of the pupillary responses dataset contains averaged pupil diameter data captured during each 10 second video interval for each of the 10 participants. There are 541 pupil diameter values for each participant and these values are averages across the fake smile stimuli and real smile stimuli categories. This aggregated dataset does not contain any missing values and in addition, it also contains gender information for each participant which will be used for our modelling. This dataset will be great for the MLP model and this dataset contains 20 samples.

The raw pupillary response dataset is very granular in nature. For each participant, we have information on their pupil diameter from both the left and right eyes, as well as which stimuli category they were viewing at the time. This raw version however, contains a number of missing values, and the number of data points for each participant varies in length. Some sequences are much longer than others, making this dataset great for the LSTM model. However, this raw version contains data for 11 participants. To keep the MLP model results and the LSTM model results comparable, data for participant p3 will be removed when building the LSTM model using this version of the data. The gender information derived from the aggregated version for each participant will be added to this version. This dataset contains 360 samples.

2.2 Aggregated Dataset Challenges and Approach

The challenge with fitting an MLP model directly to the aggregated dataset is that we have 542 input neurons but only 20 samples from 10 participants. If we treat the 541 pupil diameter values and 1 gender value as input features, we have 542 input neurons. Dimensionality reduction techniques will not help to decrease the input size space since there are only 2 core variables in this dataset, but the number of inputs is very large. An advanced data augmentation technique can be considered by fitting a Gaussian Process [10] to the real and fake classes separately to generate new observations. However for the purposes of this paper, a feature selection approach will be used where 6 handcrafted features will be created based on the findings in the original paper for the dataset. This will help to dramatically reduce the number of inputs while strategically feeding the neural network with more informative features.

2.3 Data Preprocessing for the Aggregated Dataset

Several data preprocessing steps need to be applied prior to training the MLP classification model. Our MLP classification model requires the classes to be predicted, i.e. the labels 'Real' and 'Fake', to be encoded as numeric values representing each category. As such, we encode real and fake labels as 1 and 0, respectively. All data related to groups of individuals will not be used for modelling, such as the average pupil sizes across all participants, all males, and all females since this information is not available from the raw dataset. The aggregated dataset is very balanced, with an even split between the amount of fake and real labels.

Table 1 below summarises the 6 handcrafted and refined features that will be used as inputs to the neural network model, as derived from the reason/interest point finding from the original paper for the dataset. The features above were refined based on visual inspection from plots in the original paper of how they were able distinguish or separate the classes.

Table 1. Manually selected features for neural network model.

Level	Feature	Interest Point in Findings
Across all	Pupillary size at 4.56s	Pupil dilation differed significantly from real and fake smiles at 4.56s
Across all	Pupillary size at 8.65s	Pupil dilation differed significantly from real and fake smiles between 8.62s and 8.67s
Males vs Females	Sex (1 for Male, 0 for Female)	Patterns differ between males and females
Males	Rate of change between 4s and 10s	Continuous pupil dilation from 4s to 10s in males for fake smiles
Males	Average of pupil sizes from 7.75s to 7.9s	Pupil diameter significantly large for fake smile stimuli in males from 7.75s and 7.93s
Females	Rate of change between 3-3.5s	Real smiles have a much steeper rate of change in pupil size from 2-4s in females

We can derive the value at t seconds through a simple formula. Since 541 values were recorded across 10 seconds, the value at t seconds can be derived by 541/10 multiplied by t to give us the index we should be looking for. After creating the 6 custom features, we check that these features indeed provide predictive power in distinguishing between the real and fake categories. For example, Figure 1 below is a plot of the pupillary sizes of each participant at 4.56s and at 8.65s against the real or fake labels.

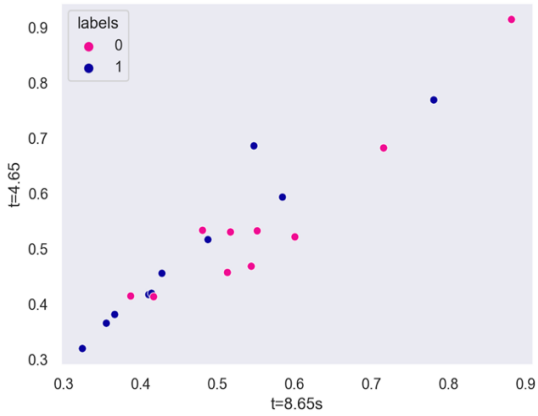

Figure 1: Relationship at 4.56s and 8.65s against Labels

The scatterplot shows that higher values of pupillary sizes at 4.65s and 8.65s corresponds closely to label = 1 (real smiles) compared to label = 0 (fake smiles) so there is a relationship which the neural network model can discover to the distinguish between the labels. Another example is shown below in Figure 2 below where a combination of gender and the rate of change at 4s and 10s could be strong predictors for a real or fake smile. For males (gender = 1), the negative values of rate of change at 4s and 10s are correlated to a real smile (class = 1) whereas for positive rate of change values, they are more correlated to a fake smile.

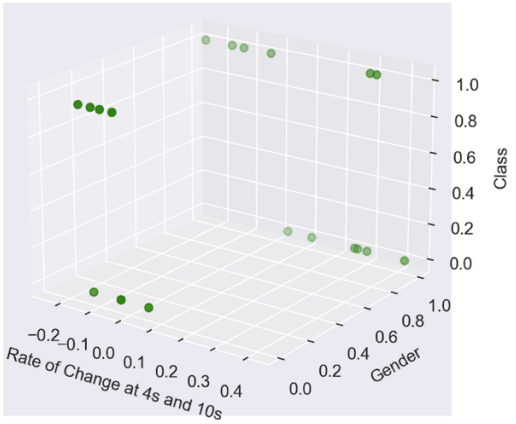

Figure 2: 3D Scatterplot of Rate of Change at 4s and 10s against Gender and Classes

Overall, the neural network model should provide some decent results given this smaller set of targeted features, even though it was all derived from the pupil size feature. Finally, the training data is normalized using the Min-Max Normalisation method [5] as a standard preprocessing step for neural networks and works well with the pupillary dataset since there are no outliers.

2.4 Training a Baseline Vanilla Neural Network

A vanilla, feedforward, fully connected MLP model will be trained and used as our baseline model for predicting real and fake smiles. This baseline model will be initialised with 6 input neurons corresponding to the 6 manually crafted input features, have 1 hidden layer, and 2 neurons in the output layer. The exact number of neurons in the hidden layer as a hyperparameter to be determined during training.

The activation functions for the hidden layer and the output layer will be the ReLU and Softmax activation functions respectively. The ReLU does not suffer from the vanishing gradient phenomena compared to the standard Sigmoid activation and it is also computationally more efficient. The Softmax activation function scales the numbers/logits into probabilities which can then use to determine the prediction made for a given input (being the maximum probability).

Since this is a classification problem, the MLP model will use the Cross-entropy Loss function and Adaptive Moment Estimation (Adam) [11] will be used as the optimiser to minimise this loss function. Adam is an adaptive learning algorithm which has faster convergence and is more reliable in reaching a global minimum, which will work well for this small dataset.

A 20% holdout test set will be reserved for testing after the optimal hyperparameters set has been confirmed. The test set needs to be designed such that information does not leak between it and the training set introducing biases. As a result, the test set will contain both the real and fake sample data for 2 participants, 1 male and 1 female, which means we will have 4 samples in total for testing. Due to the finite sample size (20), we are sacrificing accuracy for reserving 4 data points for testing, however, this is the smallest test size we can have to keep our test results reflect our model performance. For example, with only 1 test sample, our accuracy will be either 0% or 100% and this is not a good estimate of actual model performance. The same scenario applies for 2 samples.

A Leave One Out Cross-validation (LOOCV) technique will be used on the remaining 80% of the data for model training and selection to assess the predictive power of the baseline model. This 80% will be split into

a training and validation set at each fold and a manual search will be run to determine the optimal set of hyperparameters.

For each hyperparameter set at each fold, the neural network will be trained on the training set and its accuracy on the evaluation set will be recorded. Accuracy is the preferred as the model evaluation measure since we only have 1 data point in the evaluation set. The optimal hyperparameter set will be chosen based on the best average accuracy score over all 16 folds. The corresponding average accuracy to this optimal hyperparameter set will be reported as the generalisation error for the final baseline model. Due to the small dataset and the potential of the model overfitting, a value lambda was also added as a hyperparameter for L2 regularisation on the model. As such, the hyperparameters and the hyperparameter search space which the LOOCV is evaluating across is shown in Table 2 below.

Table 2. Hyperparameters for MLP and their search spaces.

Hyperparameter	Search Space
Lambda L2 Regularisation	[0, 0.01, 0.05, 0.1, 0.2, 0.4]
Neurons in Hidden Layer	[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
Epochs	[100, 150, 200]
Learning Rate	[0.001, 0.1, 0.05, 0.1]

2.5 Neural Network Pruning Technique of the Hidden Neurons

Pruning of neural networks can help to decrease model complexity, making it more interpretable and desirable. In 1995, Gedeon [8] compared model performances for 2 neural network pruning techniques which assesses the functionality of each hidden neuron. Gedeon concluded that the distinctiveness angle calculation by using the activation vectors from the hidden layer were better than the distinctiveness calculation by using the static weight vectors for the given image compression task.

Motivated by Gedeon's work, this paper applies the hidden neuron pruning technique on the neuron output activation vectors to determine a neuron's distinctiveness and whether it should be pruned. Gedeon et. al. proposed that the distinctiveness property [12] of hidden neurons is based on the neuron output activation vector over the patterns in the training set. For each hidden neuron, calculate the output activation values of the neuron across all training samples to construct a vector. This gives a matrix of size $m \times n$ where m is the number of training samples and n is the number of hidden units. Each column of the matrix is normalised to be between [-0.5, 0.5] before calculating the angle between the hidden neurons (as the indicator of distinctiveness). Hidden neurons that are similar have angles less than 15° and complementary hidden neurons have angles more than 165°. In both cases, the decision is to remove one from these pairs. In the removal process, one neuron is selected and removed by manually setting its weight to zero (mimicking the direct removal of the neuron from the network), and the weight vector of the neuron removed is added to the weight vector of the other neuron.

In this paper, this pruning technique will be used on the hidden neurons in the baseline model and a comparison on model performance with and without pruning will be analysed.

2.6 Data Preprocessing for the Raw Dataset

Since the raw dataset is split across 2 Excel files and multiple worksheets, the first step was organizing the sequential data into the same tabular format. The left and right eye pupil data are stored in different files. In each file, there are multiple worksheets corresponding to the code of the stimuli materials which was shown to each participant. Stimuli codes beginning with 'A' are non-genuine smile stimuli and stimuli codes beginning with 'L' or 'H' represent genuine smile stimuli. In each worksheet, the participant id and the recorded pupil diameter sizes are listed with varying lengths and missing values.

In order to best organize the pupil diameter information, we can create columns to store information for the participant id, participant gender, which eye the data comes from, the stimulus material code, the pupil diameter value recorded, and the ground truth label (real or fake). For modelling, only the pupil diameter and

gender features will be used as inputs. The participant id, which eye the pupil information was recorded from, and the stimulus material code will not be used as inputs into the LSTM model.

Several data preprocessing steps also need to be applied prior to training the LSTM classification model. The ground truth labels 'real' or 'fake' are encoded as numeric values 1 and 0 respectively, representing each category. The real and fake classes are quite balanced, with 47 % in the real class and 53% in the fake class and there are 370 sequences in total.

Missing values will be dealt with based on 2 scenarios. In the first scenario, if the sequence starts with or ends with missing values, these missing values will be removed since they do not provide any useful information for modelling purposes. This means that each sequence begins or ends with an observed value. In the second scenario, missing values within a sequence will be interpolation using linear interpolation. The assumption is that in the simplest case, there is a gradual increase or decrease in pupil sizes between 2 known points and there are no stochastic or drastic changes affected by the environment (environment was controlled in the study).

In contrast to the feature selection step for the aggregated dataset, we will not perform feature selection for the raw dataset due to the LSTM model as explained in the next section. Since each sequence length can be very different, it is extremely difficult to pinpoint the correct pupil dilation at a timestep t with confidence, using the same equation we used for the aggregated dataset. In essence, the LSTM will accept 2 input neurons due to its architecture, being the pupil dilation size at a particular timestep and the gender of the participant.

The Min-Max Normalisation will also be applied to the training data, with the same normalisation factor applied to the testing data before predictions are made.

2.7 LSTM Model

The Recurrent Neural Networks (RNN) [13] is a popular neural network architecture for modelling sequential data due to its performance and application to important tasks such as speech recognition [14], time series prediction, and image captioning. RNNs contain connections between input units, allowing them to capture dependences on inputs through time. This connection feeds the hidden state of the previous input into the current input and are effective in handling sequential data with varying lengths such as the raw pupillary response dataset. Figure 3 [15] below shows the general structure of an RNN on the left hand side which can be unfolded through time and represented as the network on the right hand side.

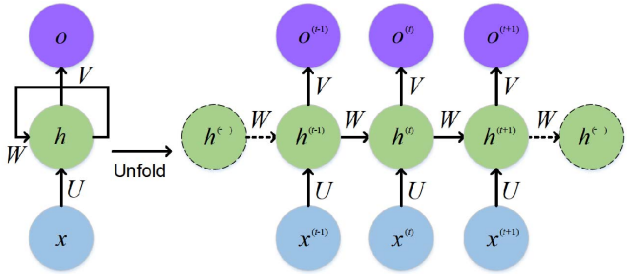

Figure 3: RNN Network Structure

However, RNNs suffer from the vanishing or exploding gradient problem in backpropagation through time [16]. For our raw dataset, sequence lengths can range from over 700 to over 1,100 so using an RNN can cause issues. LSTMs are an implementation of RNNs which overcomes this issue. LSTMs house a memory cell, shown in Figure 4 [17] below, controlling 3 gates that decide when and what type information in the network should be fed into the current hidden unit.

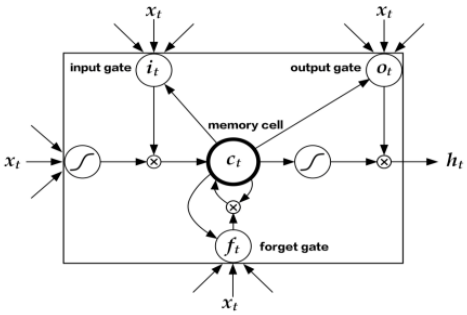

Figure 4: LSTM Memory Cell

We will build a simple LSTM model with 1 hidden layer using the raw data in understanding if sequential or temporal information can help us better predict a real smile from a fake one. As with the MLP case, the LSTM classification model will use the Cross-entropy Loss function and the optimizer Adam. Although the classes of real and fake samples are quite balanced, care needs to be taken when splitting the dataset into a training, validation and testing set. We need to ensure that we do not leak information between the different sets of splits. In our case, we will split the data into 70% training, 20% validation, and 10% testing based on participant level information. This ensures that data samples for a participant such as their left and right pupil information stay in the same split. For training, we will use a batch size of 1, since all the sequences are different lengths.

The evaluation set will be used to optimize over the hyperparameter set for the LSTM model. The optimal hyperparameter set will be chosen based on the best accuracy score. The final LSTM model will be trained using the optimal hyperparameter set on the full training set (90% of the data) and evaluated against the test set. The hyperparameters and the hyperparameter search space used in this paper is shown in Table 2 below.

Table 3. Hyperparameters for LSTM and their search spaces.

Hyperparameter	Search Space
Lambda L2 Regularisation	[0, 0.1, 0.2, 0.3]
Neurons in Hidden Layer	[5, 10, 50, 75, 100, 150, 200]
Epochs	[10, 50, 100]
Learning Rate	[0.01, 0.05, 0.1, 0.2, 0.3]

3 Results and Discussions

3.1 Baseline Neural Network Model

After running LOOCV on the 1,368 hyperparameter combinations on the baseline model, the best set of hyperparameters found was epoch = 100, learning rate = 0.1, number of hidden neurons = 7, lambda = 0.01 with an average validation accuracy score of 62%. It seems that the baseline model benefits from a very small amount of regularisation. The training loss per epoch is shown in the line plot in Figure 5 below.

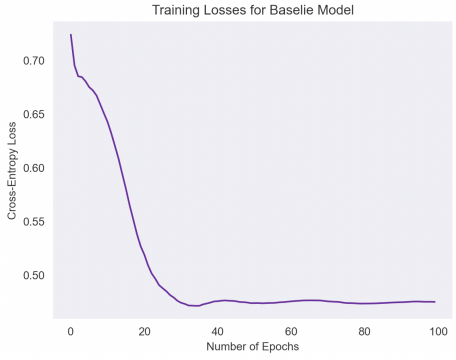

Figure 5: Training Loss of Baseline Model

As expected, the training loss for the baseline model decreases as the number of epochs increase which means the model is training well. The final training loss for the baseline model was to be 0.475 and the training accuracy was 81.25%. However, the test accuracy for this model was 50%. If we have a look closer in what this baseline model is outputting in its predictions for the 4 test cases, the baseline model was making all predictions for class 1 (real smile) with over 60% probability in all cases.

3.2 Pruned Neural Network Model

2 iterations of network pruning were performed on the baseline model to reduce the model complexity. Where a decision had to be made on which hidden neuron to keep, the earlier numbered neuron was removed by forcing its weights to zero. The results are summarised in Table 4 below.

Table 4. Distinctiveness Neuron Pruning Iteration Results.

Iteration Number	Smallest Angle	Neuron Pair Decision	Pruned Model Test Accuracy
1	1.16°	Remove 2, Keep 3	50%
2	1.21°	Remove 4, Keep 5	25%

After pruning 2 hidden neurons from a model of 7 hidden neurons, we found that the test accuracy of the neural network decreased from 50% to 25%. This means that we should only remove one neuron from the model in order to maintain our accuracy in predictions. By performing distinctiveness pruning on the activation output vectors, we can reduce the number of hidden neurons by 1 neuron (14%), thereby simplifying the model to reduce redundant neurons.

If we take a look closer at what this pruned model is outputting in its predictions on the same 4 test cases, this pruned model actually predicted 2 real smiles and 2 fake smiles, which is better than predicting all of the same class (baseline model). However, 2 of these predictions were very borderline cases, meaning the network has become more uncertain about these cases after removing a neuron.

Neural networks are capable of modelling both linear and complex, nonlinear relationships among variables [18]. However, even though the features we generated look promising in detecting real smiles, it seems that there is no complex relationship between pupil size and gender which can help us discriminate real from fake smiles.

3.3 LSTM Model

For the LSTM model, we experimented with including and excluding gender information. Although in the original paper for the data, it was found that gender could provide predictive power for our classification task, we found that the gender information did not provide any advantages for the LSTM model. The evaluation set accuracy was almost identical for the same hyperparameter sets. This may be because gender is a feature which does not change its value over time, so gender does not provide any additional predictive support for the LSTM model. For example, the LSTM model might learn to always forget the historical patterns for gender since it will always receive it in its current input, or vice-versa. As such, the final model after hyperparameter testing will only consider the pupil size as the input.

After running different hyperparameter combinations on the LSTM model, the best set of hyperparameters found from our hyperparameter search space was epoch = 50, learning rate = 0.2, number of hidden neurons = 50, lambda = 0.2 and the evaluation accuracy was 52%. It seems that the LSTM model benefits from some regularisation. The final model was trained on 90% of the data and the final model had a testing accuracy score of 50%. It turns out that the LSTM model was making all its predictions for class 0 (fake smile) on all 26 test data with over 80% probability.

3.4 Discussion

In all cases, the test sets were very balanced in terms of real and fake samples. However, for the baseline, pruned, and LSTM neural network models, the accuracy of the models on the test set did not surpass 50%. In other words, the neural network models performance is as good as anyone random guess by chance. What is very surprising is that the baseline model was predicting that everything was a real smile whereas the LSTM model was predicting everything was a fake smile. The baseline model became more uncertain in its predictions after one hidden neuron was removed, despite holding the same accuracy score on the test set. Although the pruned model would generally be more preferred due to its simplicity, it is very difficult to conclude which model was superior due to the finite number of samples available for all models. We also saw that although we were able to prune the model and maintain the same test accuracy, the pruned model became more uncertain in its predictions. In addition, with increased data samples, these models will be able to make better quality predictions and this is especially true for the deep learning model [19]. Another popular method is transfer learning [20], which helps deep learning models overcome the issue of finite data.

Transformer [21] is another popular type of RNN model and is state of the art when it comes to applications on sequential data such as natural language processing or time series tasks. The advantage of modelling with the transformer model is that it uses multi-head attention – a mechanism which is able to learn contextual information. The transformer model could be a good application to this dataset. From the original paper for our dataset, we know that there are different points in time for pupil dilation which were statistically significant and we used these points in time to handcraft our features for the baseline model. The transformer model could learn to pick up these contexts in the sequence to better predict a real smile from a fake smile.

Even though gender influences pupil size in the study from the original paper for this dataset, we found that static features such as gender had no impact on the LSTM model's predictive power. We also found that there was no strong linear or non-linear relationship between gender and pupil size that could help us detect real from fake smiles. This suggests that we cannot use pupil size alone to detect genuine smiles and that we may need other features such as EGG data or we may need to enrich our dataset with another feature that is to complimentary with pupil size.

4 Conclusion and Future Work

Using the findings from a study of the relationship between pupillary size and fake or real smiles, this paper implemented 3 neural network models in attempting to detect real smiles using pupillary data from 10 participants. The first model was referred to as the baseline model which is a 1 hidden layer, MLP model with optimised hyperparameters. The second model is a pruned version of the baseline model using angle distinctiveness based on activation vectors of the hidden neurons. Using this pruning technique, we were able to prune 1 hidden neuron from the baseline model before sacrificing accuracy on the test set. The final model was a 1 hidden layer LSTM model with optimised hyperparameters. In all cases, these models all achieved 50% accuracy on the corresponding test sets. As a result, there are no advantages in using a more complex LSTM model to capture memory of earlier sequential inputs for this dataset.

We found that the models performing at the same accuracy as random guessing. This could be due to a few factors. One of these factors could be related to the finite sample size, as we know that neural networks, especially deep neural networks, benefit from large amounts of data. Another factor worth mentioning is that the pupillary size and gender feature combinations may not provide enough predictive power in detecting real smiles compared to EEG signals which gave higher performances[4]. We can consider applying transfer learning or collecting more data in these cases. In our case, we also found that there was a trade-off between model prediction confidence and model complexity when we pruned our baseline model. This could be the influence of the finite sample size or indicate a more complex model with more hidden layers is required.

In future work we can consider more advanced RNN techniques such as Transformer models, or improve current models by adding additional hidden layers. In the case of the latter, the distinctive network pruning method can still be applied. We can also consider ensemble methods such as bagging or boosting [3] with other machine learning models which could help to improve the performance of the model and generate

better insight. Performing automated grid search or random search with a larger hyperparameter search space could also give better modelling results.

5 References

1. Kołakowska A., Landowska A., Szwoch M., Szwoch W., Wróbel M.R. (2014) Emotion Recognition and Its Applications. In: Hippe Z., Kulikowski J., Mroczek T., Wtorek J. (eds) Human-Computer Systems Interaction: Backgrounds and Applications 3. Advances in Intelligent Systems and Computing, vol 300. Springer, Cham. https://doi.org/10.1007/978-3-319-08491-6_5
2. Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. Journal of Personality and Social Psychology, 58, 342–353.
3. Valstar, M. F., Gunes H., Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles using Geometric Features, Proceedings of ACM International Conference on Multimodal Interfaces (Nagoya, November 2007), 38-45.
4. M. Moussa, U. Tariq, H. Al-Nashash and F. Al-Shargie, "Discerning Genuine and Acted Smiles Using Neural Networks," 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2019, pp. 1-4, doi: 10.1109/ISSPIT47144.2019.9001848.
5. Basheer, I. and M. Hajmeer. “Artificial neural networks: fundamentals, computing, design, and application.” Journal of microbiological methods 43 1 (2000): 3-31 .
6. Hossain, M. Z., Gedeon, T., Sankaranarayana, R., Apthorp, D., & Dawel, A. (2016). Pupillary responses of Asian observers in discriminating real from fake smiles: A preliminary study. In Measuring Behavior (pp. 170-176).
7. Zhang, W., Du, Y., Yoshida, T., Wang, Q., DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network, Information Processing & Management, Volume 54, Issue 4, 2018, Pages 576-592, ISSN 0306-4573 (2018)
8. T. D. Gedeon, "Indicators of hidden neuron functionality: the weight matrix versus neuron behaviour," Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, 1995, pp. 26-29, doi: 10.1109/ANNES.1995.499431.
9. Bouktif S, Fiaz A, Ouni A, Serhani MA. Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches †. Energies. 2018; 11(7):1636. https://doi.org/10.3390/en11071636
10. M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In Artificial Intelligence and Statistics, pages 567–574, 2009. Cited on pages 2, 4, 16.
11. Jais, I.; Ismail, A.; Nisa, S. Adam Optimization Algorithm for Wide and Deep Neural Network. Knowl. Eng. Data Sci. 2019, 2, 41.
12. Gedeon, TD, Harris, D, “Network Reduction Techniques,” Proc. Int. Conf. on Neural Networks Methodologies and Applications, AMSE, San Diego, vol. 2, pp. 25-34, 1991.
13. Alex Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, Physica D: Nonlinear Phenomena, Volume 404, 2020, 132306, ISSN 0167-2789, https://doi.org/10.1016/j.physd.2019.132306.
14. Graves, Alex. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013.
15. Feng, W., Guan, N., Li, Y., Zhang, X., Luo, Z. (2017). Audio visual speech recognition with multimodal recurrent neural networks. 681-688. 10.1109/IJCNN.2017.7965918.
16. S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, “Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies,” 2011.
17. Plested, J. (2021) ANU COMP8420 Neural Networks, Deep Learning and Bio-inspired Computing. Sequence Learning.
18. Curry, B., Morgan, P.H. Neural networks, linear functions and neglected non-linearity. Computational Management Science 1, 15–29 (2003). https://doi.org/10.1007/s10287-003-0003-4
19. Hao X, Zhang G, Ma S. Deep learning. IJSC. 2016;10:417-439.
20. L. Torrey and J. Shavlik. 2009. Transfer learning. In E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, editors, Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global.
21. P. T. M. Vaessen, "Transformer model for high frequencies," in IEEE Transactions on Power Delivery, vol. 3, no. 4, pp. 1761-1768, Oct. 1988, doi: 10.1109/61.193982.