Variable thresholds for the classification of genuine or posed anger

Will Orr
U6655462@anu.edu.au

Abstract

Other's emotions are not just registered consciously but can also produce non-conscious physiological responses in observers. This study investigates the effectiveness of a variable threshold for a recurrent LSTM neural network trained to classify genuine and posed displays of anger from the involuntary variation in an observers' pupil size. Models were trained and tested on three different arrangements of the data to investigate the impact reserving different types of data on model accuracy. A variable threshold can be useful in order to reduce or even eliminate classification errors, as well as improve accuracy. Though the utility of thresholds that eliminate classification errors differed between data arrangements, varying the classification threshold led to an improved accuracy of the model, achieving an overall accuracy of 85.5%. While this was a lower classification accuracy than previous studies with this data, this model outperformed observer's conscious determinations of the sincerity of anger. Variable classification threshold could therefore be a useful addition for future emotion classification research.

1. Introduction

Emotions are not only interpreted by the human brain; rather, the human body too can produce distinct involuntary physiological reactions when interpreting the emotional nature and sincerity of another's actions (Hossain et al., 2016; Chen et al., 2017). This research seeks to consolidate upon previous investigations into classifying real and fabricated displays of anger from the pupillary responses of an observer, through the creation of a binary neural network classifier (Chen et al., 2017). While unconscious physiological responses to anger may be better predictors of emotional veracity than individuals' own conscious determinations of emotion, uncertainty still exists, and neural net classification errors still occur (Chen et al., 2017). This research hence seeks to explore the applicability of threshold variability for the classification of genuine and posed displays of anger in order to minimize False Positive and False Negative errors and achieve superior accuracy.

As noted in Kogan (1991), the output node of a binary neural network classifier cannot be viewed as a confidence factor for the classification. As such, the threshold for a positive classification is not fixed at the traditional value of 0.5 but rather, can be varied (Kogan 1991; Milne et al., 1995). Informed by Kogan (1991), Milne et al. (1995) implement a variable classification threshold for identifying pixels containing dry sclerophyll forest from satellite imagery and aerial photographs.

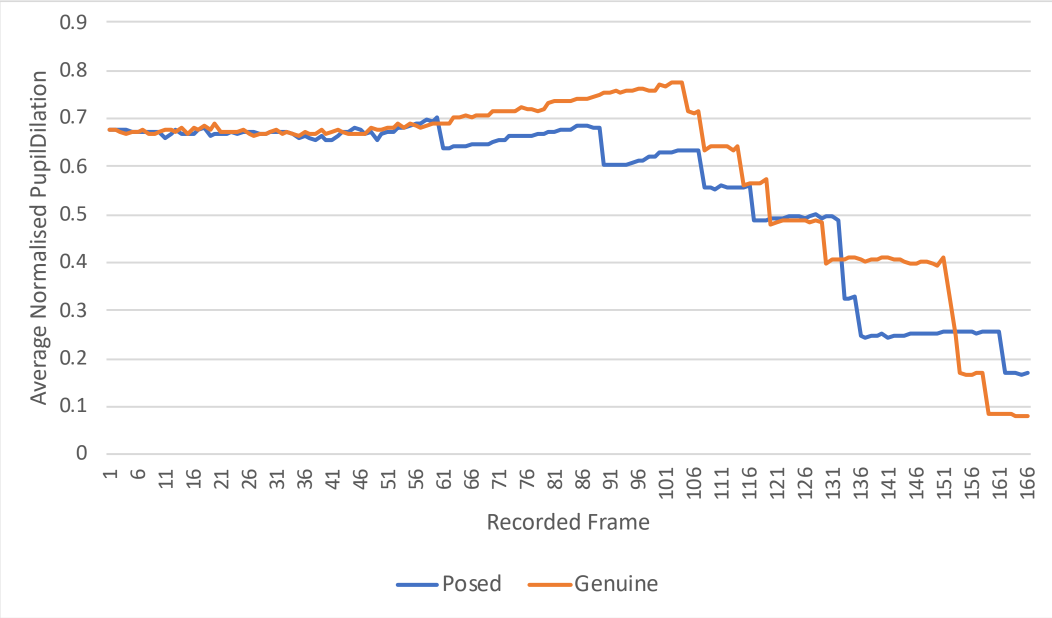
A variable threshold for classification is particularly useful when the costs of False Positives and False Negatives are not the same (Corbett-Davies et al., 2017). In medical classification, a False Positive result leads to a strain on medical resources and induced anxiety on the patient and their family; a False Negative classification, on the other hand, can cost valuable treatment time and potentially be fatal (Corbett-Davies et al., 2017). Adjusting the threshold for classification and identifying the points at which False Positive and False Negative classifications are minimized, as done by Milne et al. (1995), can be a useful means of balancing these costs. This research will thus implement the techniques outlined by Milne et al. (1995) on an entirely different domain: the classification of emotional sincerity.

1.1 The Data

The data used for this research is drawn from Chen et al.'s (2017) investigation into physiological responses of genuine and posed states of anger. In their experiment, 20 participants watched video clips of real and acted presentations of anger, 10 of which were genuine displays of anger from news recordings or documentaries and 10 were acted within movies or tv shows. Participants' right and left eye pupillary dilation response to viewing these clips were recorded at regular intervals for each of the 20 videos.

To make use of this data, the average pupil dilation of each time interval for each participant was taken, for a total of 387 observations. These values were then normalised between the maximum and minimum pupil dilation exhibited for each participant over the course of the 20 videos. As some videos are longer than others, videos shorter than 186 timesteps were padded with zeros at the end of the sequence. Finally, each video was designated a target value of either 1 for genuine displays or 0 for acted presentations of anger. The following figure displays the average normalised pupil response to viewing videos of both genuine and posed anger over the course of each video for each participant.

Figure 1: Normalised pupillary response to real and posed displays of anger over time



Following data pre-processing, the data were split into training, validation and testing sets. How these data are split however, inherently embodies assumptions about its nature and general applicability. As such, three different training, validation and test data sets were made from the data, each with their own set of assumptions. These three different cases will be compared and used to train, validate and test the models.

To construct the first datasets, video and participant information was ignored (referred to as the randomised datasets). The records were shuffled with 15% was randomly assigned to a validation set and a further 15% to the test dataset. The remaining 70% of the records were used to train the model. By not discriminating by video or participant, this dataset split assumes that there is little difference between the physiological responses of individuals or responses to the individual videos. It therefore should not matter if the model has already been trained on an observation of a particular video or by a particular participant.

The second datasets discriminated by video (referred to as video datasets). Each video may incite the same physiological response in participants due to the course of tension unique to the video itself.

Furthermore, with different length of padding, the model may learn these characteristics rather than the pupillary responses. Videos were randomly selected and reserved for the validation and test datasets. To maintain similar class distribution, each validation and testing sets contained responses to one genuine and one posed video, as well as half of the responses to a video of each class (approximately equating to 15% of the data each). The pupillary responses to the remaining 14 videos were used to train the model. The model will therefore be tested on responses to videos that have not been seen in training.

The final datasets discriminated by participant (referred to as participant datasets). The validation and testing datasets were both randomly assigned three participants, roughly equating to 15% of the overall data each. The remaining 14 participants comprised the training dataset. This assumes that participants may have unique characteristics when viewing genuine or posed displays of anger that could be learnt by a model. These datasets aim to address this concern by testing the model on participant information unseen by the model.

2. Methods:

To classify pupil responses of participants as observations of either genuine or posed displays of anger, a Long Short-term Memory (LSTM) artificial recurrent neural network is constructed with single input and output neurons. As this network will be a binary classifier, binary cross entropy loss is selected as the loss criterion. The Adaptive Moment Estimation (Adam) optimiser function is selected has been successfully used for robust optimisation of previous LSTM models (Kong et al., 2017). The final prediction must be either 0 or 1; hence, the output of the final layer must be a value between 0 and 1. As such, the sigmoid activation function is selected for the final output neuron.

2.1 Hyper-parameter selection:

In order to select the optimal hyper-parameters, a number of tests were conducted. To identify the most optimal settings, the number of hidden neurons, learning rate, number of epochs, number of layers and the dropout rate were varied and the effects of these settings on the three validation datasets were compared. To maintain consistency, each test was conducted with the default classification threshold of 0.5 and a batch size of 32. It must be noted that this was an iterative process, moving back and forth between these hyperparameter tests to identify the most optimal balance of each.

2.1.1 Hidden Neurons:

To identify the optimal number of LSTM hidden neurons for training, single hidden layer LSTM models were constructed with the number of neurons ranging from 5 to 25. For this test, the learning rate was kept consistent at 0.001 and the number of epochs at 200.

2.1.2 Learning Rate:

To investigate the effect of varying the learning rate on training, rates between 0.05 and 0.0001 were tested. These tests were conducted with a single layer model with 20 hidden LSTM units over 100 epochs

2.1.3 Epochs:

To optimise the number of epochs, a graphical approach was taken. By comparing the training and validation loss graphs, the optimal number of epochs to avoid over-training was identified. These models were trained with a single hidden layer LSTM model with 20 hidden neurons, with a learning rate of 0.0005.

2.1.4 Number of Layers:

To test the impact on the number of layers on training, models of 1, 2 and 3 hidden layers were constructed. Each of these models contained 20 hidden neurons and had a learning rate of 0.0005 and was trained over 120 epochs.

2.1.5 Dropout Rate:

Overfitting has been a reported issue with LSTM recurrent neural networks (Merity et al. 2017) and indeed was a problem in a previous iteration of this research (Assignment 1). To combat this issue, dropout rates for the LSTM layer between 0 and 0.5 were tested using a two hidden-layer LSTM model with 20 hidden neurons and a learning rate of 0.0005, over 120 epochs.

2.2 Classification threshold

After selecting the most optimal hyper-parameters, the threshold to which an output is classified as a display of genuine anger will be explored. There are two aims in varying the classification threshold of this recurrent neural network. Firstly, the thresholds at which false positive and false negative classifications are minimised will be identified for each of the three test datasets. And secondly, the associated accuracy for each classification threshold will be identified, thus optimising the classification accuracy of the model. Following Milne et al. (1995), the threshold of the final layer output for positive classification is varied by 0.1 increments. For each step, the number of False Positive and False Negative classifications in each test dataset are recorded, as well as the mean test accuracy across these three cases. The thresholds that result in no False Positive and False Negative classifications for each dataset will hence be identified. These results will be compared with the results obtained from a fully-connected artificial neural network with one hidden layer of 10 neurons, a learning rate of 0.05. This model was trained over 200 epochs using aggregate data such as the mean and standard deviation of pupil dilation for each observation, as well as principle component analysis.

3. Results

3.1 Hyperparameter selection

Conducting the tests outlined above, the following results were obtained:

3.1.1 Hidden Neurons:

Hidden Neurons	Randomised Validation Accuracy (%)	Video Validation Accuracy (%)	Participant Validation Accuracy (%)	Mean Training Accuracy (%)	Mean Validation Accuracy (%)	Difference Training - Validation
5	57.6	19.6	64.4	68.9	47.2	21.7
10	54.2	53.6	52.5	59.6	53.5	6.2
15	57.6	48.2	76.3	65.0	60.7	4.3
20	79.7	71.4	55.9	72.0	69.0	3.0
25	54.2	44.6	55.9	66.9	51.6	15.3

As the number of LSTM neurons increased, the accuracy of each validation set increased before dropping off when too many neurons were added. This drop in accuracy could be considered a sign of the model overfitting, as it ‘memorises’ the training data, but does not work as well on the validation set. The model accuracy on randomised data and unseen video validation datasets both peaked with a hidden layer of 20 neurons, before decreasing dramatically with 25 hidden neurons. However, the accuracy of the model on the unseen participant validation set peaks with a hidden layer of 15 neurons before decreasing with 20 and 25 hidden neurons. This suggests that different methods of dividing data for training and validation may influence how well a model may learn the available data and how easily it may over-train.

While the mean training accuracy fluctuated with the addition of more hidden neurons, the difference between the mean training accuracy and the mean validation accuracy steadily declined until 20 hidden neurons, before increasing again dramatically at 25 hidden neurons. The large difference between average training and validation accuracy with 25 hidden neurons is likely an indication that the model is overtrained on the training data, while the large difference between average training and validation accuracy at only 5 hidden neurons suggests that the model is undertrained and is not yet familiar with the data.

While 20 hidden neurons were not optimal for every validation dataset, it produced the greatest average training accuracy and test accuracy of all tested combinations, as well as the smallest difference between average training and validation accuracies and will therefore be used in future trials.

3.1.2 Learning Rate:

To identify the optimal learning rate for training, the following results were obtained. These recurrent LSTM models were trained over 100 epochs with 20 hidden neurons.

Learning Rate	Randomised Validation Accuracy (%)	Video Validation Accuracy (%)	Participant Validation Accuracy (%)	Mean Training Accuracy (%)	Mean Validation Accuracy (%)	Difference Training - Validation
0.05	54.2	16.1	50.9	62.2	40.4	21.8
0.01	78.0	48.2	55.9	65.2	60.7	4.5
0.005	79.7	66.1	55.9	73.2	67.2	6.0
0.001	78.0	71.4	54.9	72.1	68.1	4.0
0.0005	72.9	71.4	83.1	80.4	75.8	4.6
0.0001	54.2	48.2	49.2	53.3	50.5	2.7

High learning rates had little benefit to either training or testing accuracy, due to a large fluctuation in neuron weights. Notably, the model validated against the unseen video observations performed particularly poorly with a learning rate of 0.05, performing well below mere random. The model validated against the randomised validation set performed well with learning rates of 0.01, 0.005 and 0.001 before decreasing in accuracy with a learning rate of 0.0005. This is in contrast to the unseen participant validation dataset which achieved poorer accuracy at learning rates 0.05, 0.01, 0.005 and 0.001 before a rapid increase at a rate of 0.0005, then falling again at a rate of 0.0001. Each model performed poorly with a learning rate of 0.0001 indicating that the model likely undertrained. A learning rate of 0.0005 was selected for future analysis as this had the largest average validation accuracy and was the only learning rate that performed well with unseen participant data.

3.1.3 Epoch

The optimal epoch for training the model was found by comparing the loss of the training model to the loss of the validation model for each epoch. The loss graph for one of models of a network with 20 hidden neurons with a learning rate of 0.0005, a batch size of 32 over 150 epochs is as follows:

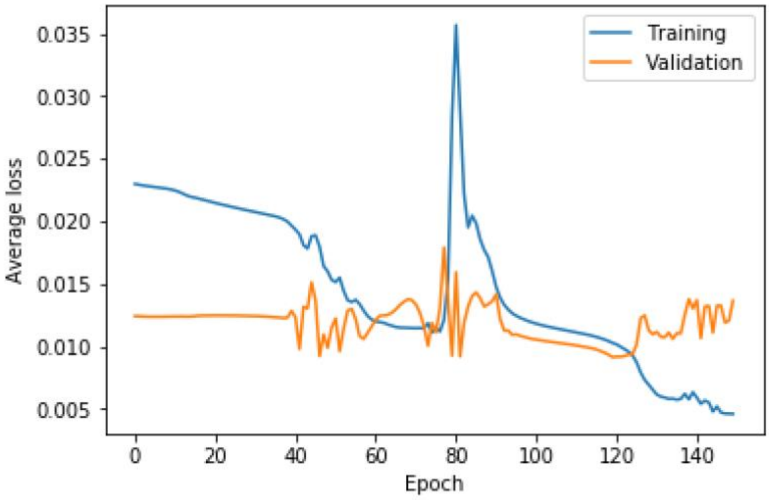


Figure 2: Training and Validation Loss for identifying optimal number of epochs

The training and validation loss diverge following the 120th epoch, suggesting that the model begins to overfit thus decreasing its effectiveness when classifying unseen data. An epoch range of 120 was therefore selected for future training to increase validation accuracy and decrease overfitting.

3.1.4 Number of Layers:

Testing the effects of the number of hidden LSTM layers, the following results were obtained:

Layers	Randomised Validation Accuracy (%)	Video Validation Accuracy (%)	Participant Validation Accuracy (%)	Mean Training Accuracy (%)	Mean Validation Accuracy (%)	Difference Training - Validation (%)
1	72.9	71.4	74.6	81.7	73.0	8.7
2	76.3	71.4	81.4	84.1	76.4	7.8
3	91.5	73.2	86.4	90.7	83.7	6.9

Increasing the number of layers led to an increase in accuracy across each of the validation sets as well as an increased training accuracy. Having three layers also led to the smallest gap between average training and validation accuracy. What is not captured by this table is the computational strain that running a model with 3 layers required on my system; python crashed on multiple occasions and if it did complete training, each epoch took on average 4.2 seconds (as compared to 0.8 seconds with 1 hidden layer and 1.4 seconds with two hidden layers). As such, due to technical limitations, I was unable to proceed with three hidden layers. This however does seem promising and could be tested in the future with better infrastructure. As such, two hidden layers were used instead as the next best alternative.

3.1.5 Dropout Rate:

To combat overfitting, dropout of LSTM units was considered. In testing the effects of dropout rates, the following results were obtained:

Dropout Rate	Randomised Validation Accuracy (%)	Video Validation Accuracy (%)	Participant Validation Accuracy (%)	Mean Training Accuracy (%)	Mean Validation Accuracy (%)	Difference Training - Validation
0	76.3	71.4	81.4	84.1	76.4	7.8
0.1	83.1	58.9	86.4	84.3	76.1	8.2

0.2	83.1	30.4	74.6	86.7	62.7	24.1
0.3	88.1	35.7	78.0	86.9	67.3	19.6
0.4	79.7	44.6	79.7	86.1	68.0	18.1
0.5	86.4	28.0	71.2	85.3	61.9	23.5

Including dropout of LSTM improved the accuracy when validated against a randomised dataset and unseen participant data, with accuracy peaking at a dropout rate of 0.3 and 0.1, respectively. Including any dropout had significant negative effects on the classification of unseen video data. This suggests that the models trained against randomised and participant-based data may have overfitted to the training data, thus making dropout a useful addition. However, the model classifying unseen video data may not need regularisation and thus, neuron dropout may lead to loss of information and classification potential. Due to the marginal higher average validation accuracy and to avoid a substantial decrease in accuracy on unseen video data, no dropout was included in future analysis.

At the conclusion of hyperparameter testing, binary classification neural network was created with two layers of 20 LSTM units, with a learning rate of 0.0005, trained over 120 epochs with a batch size of 32. Throughout this hyper-parameter tuning process, the difficulty of optimising for three different validation sets was emphasised – while some validation sets may respond well to some settings, they may have the opposite effect for other validation sets that embody different assumptions.

3.2 Classification Threshold

In testing the impact of varying the classification threshold on binary classification, the following results were obtained:

3.2.1 Minimising Errors:

Classification Threshold	**Randomised Test**		**Video Test**		**Participant Test**	
	FP	**FN**	**FP**	**FN**	**FP**	**FN**
0.2	25	0	27	0	11	0
0.3	18	0	12	0	7	1
0.4	13	0	10	0	4	4
0.5	10	2	5	0	4	7
0.6	9	2	3	0	1	10
0.7	6	5	2	0	0	13
0.8	4	7	1	0	0	13
0.9	0	10	0	1	0	16

As the threshold for positive classification increased, the number of False Positive classifications decreased while the number of False Negative classifications increased for each of the test datasets. For the randomised test set, no False Negative classifications were observed for thresholds of 0.2, 0.3 or 0.4, while remaining low at thresholds of 0.5 and 0.6. At a threshold of 0.9, no False Positive classifications were observed. For the unseen video test set, False Positive classifications decreased as classification threshold increased, with little effect on False Negative classifications which remained at zero until a threshold of 0.9. The unseen participant test set achieved a lower threshold for no False Positive observations than the other test sets; all classifications of genuine anger (positive classification) at the threshold of 0.7 were in fact genuine. False Negative errors were eliminated with a threshold of 0.2, much lower than the randomised and unseen video test set.

Each of these test sets identified different thresholds for achieving no erroneous classifications: 0.4 and 0.9 for the randomised test set, 0.8 and 0.9 for the unseen video test set, and 0.2 and 0.7 for the unseen participant test set. This suggests that there is no one size fits all approach for reducing erroneous classifications as these thresholds depend on the data that the model was trained on, and the nature of the test datasets. Nevertheless, across each of the three test sets, no False Negative classifications were observed with a threshold of 0.2 while False Positive classifications were minimised at a threshold of 0.9. While these thresholds for eliminating False Negative and False Positive classifications are less useful than the 0.5 and 0.7 thresholds identified by Milne et al. (1995) in classifying dry sclerophyll forest, these limits are more useful than those identified by the fully connected model of 0.1 and 1. This is because extreme classification thresholds are often trivial as they regard most records as either positive or negative class despite the input data, thus reducing the capacity for correct classification.

3.2.2 Maximising Accuracy:

Varying classification threshold also affected the overall accuracy of the models.

Classification Threshold	Randomised Test Accuracy	Video Test Accuracy	Participant Test Accuracy	Total		
				Accuracy	Precision	Recall
0.2	56.14	52.63	81.03	63.3	57.3	100.0
0.3	68.4	78.98	86.21	77.9	68.4	98.9
0.4	77.19	82.46	86.21	82.0	73.9	95.6
0.5	78.95	91.23	81.03	83.7	78.7	89.1
0.6	80.7	94.73	81.03	85.5	84.6	85.7
0.7	80.7	96.49	77.59	84.9	88.8	77.6
0.8	80.2	98.25	77.59	85.4	91.5	74.4
0.9	82.46	98.25	72.41	84.4	100.0	65.2

The accuracy of the models on their respective test sets followed the distribution of errors identified above. Classification on the unseen participant data achieved highest test accuracy at thresholds 0.3 and 0.4 while the randomised test set and unseen dataset achieved highest accuracy at higher thresholds such as 0.8 and 0.9. This indicated that the models were much better at correctly predicting responses to genuine presentations of anger in the randomised test set and the unseen video test set as correct positive classifications were still achieved for higher thresholds, while they were better at correctly classifying responses to posed presentations of anger in the unseen participant dataset.

The unseen video dataset particularly benefitted from increasing classification threshold, almost achieving perfect accuracy at thresholds 0.8 and 0.9. This is surprising given the relatively lower accuracy achieved by the unseen validation dataset when tuning hyperparameters. This suggests that the unseen videos in this test set may closer resemble the data used for training than those in the validation set.

Accuracy over the three test sets was maximised at a threshold of 0.6, however comparable accuracy was also achieved at thresholds 0.7, 0.8 and 0.9. At a threshold of 0.6, 85.5% of records across the three datasets were classified correctly, with comparable precision and recall values (84.6% and 85.7%, respectively). This is a marginal improvement from accuracy at the traditional default threshold of 0.5, which achieved 83.7% accuracy across the three datasets. While total accuracy remained high at thresholds 0.7, 0.8 and 0.9 as model precision increased, this came with a trade-off of decreased recall.

These benefits of variable threshold for neural networks have been recognised elsewhere. Pendharkar (2004) conceptualises classification threshold as not a static parameter, but rather a value to be learnt through the training process in order to maximise a systems classification accuracy, identifying improved accuracy in both training and evaluation compared to traditional back-propagated artificial neural networks with a 0.5 classification threshold. More recently, Pendharkar (2008) consolidates this work by creating learnable threshold that is sensitive to the marginal costs of false positive and false negative classifications. Both of these accounts highlight the benefits of a variable classification threshold as outlined in Milne et al. (1995), and the increased performance accuracy achieved here.

It must be noted that the performance accuracy achieved here is below that achieved in similar classification problems. Chen et al. (2017) created a classifier trained on the present data set of pupillary responses to genuine and posed displays of anger, achieving a 95% classification accuracy. These results are similar to the neural net classification of real and fake smiles from observers' galvanic skin responses, with a 96.5% classification accuracy (Hossain et al., 2016). The maximum achieved classification accuracy in the present research of 85.5% therefore sits below these comparable studies. Nevertheless, this accuracy still exceeds the 60% accuracy of conscious, verbal responses of participants in determining sincerity of anger (Chen et al., 2017). Furthermore, the present recurrent LSTM neural network trained on timeseries data of pupillary dilation achieved a higher accuracy than the dense artificial neural network that was trained on aggregate data and principle components, which correctly classified 77.9% of records.

4. Conclusion and Future Work

This research has identified the capacity for binary neural network classifier to discriminate between genuine and posed accounts of anger from the unconscious pupillary responses of an observer. This was completed with a recurrent LSTM neural network with two layers of 18 units, trained with a learning rate of 0.0005, a batch size of 32 over 120 epochs. A variable classification threshold was used in order to minimise classification errors and improve the models' overall accuracy. Models trained and tested on three different divisions of the data, a set comprised of randomised participants and videos combinations, a set comprised of pupil responses of all participants to videos unseen by the model in training, and a set comprised of pupil responses to all of the videos by participants whose data was unseen in training. While errors in each of these test sets were eliminated at different thresholds, at the thresholds of 0.2 and 0.9, no false negative or false positive observations were made, respectively. Increasing the classification threshold to 0.6 led to an increased total classification accuracy of 85.5%. While this exceeds the accuracy of conscious choice by participants (60%) as well as the accuracy obtained by a fully connected neural network trained on aggregate data for each observation (77.9%), it fails to match the accuracy achieved with previous applications of this data (95%) (Chen et al., 2017).

The divisions of data had a significant impact on the outcomes of hyperparameter testing and the effectiveness of variable classification thresholds. While a composite model was constructed as a means of accounting for different circumstances to achieve optimal outcomes across these three cases, their differences highlight impact of splitting data into training, validation and test sets and the assumptions that are embodied in these actions. Despite my best efforts, each of these models still may have had data about specific videos or participants leak into the validation and testing datasets. As such, further testing of this model should be conducted with fresh participants and different videos of displays of anger.

This project was also limited by technical infrastructure. While a neural network of three hidden layers performed well on all datasets during hyperparameter testing, it was not feasible for future testing due to the technical limitations of the computer used for training. Future research with stronger computational power could further investigate the effectiveness of additional hidden layers on model accuracy.

There is also scope to extend this work through the implementation of a classification threshold that is learnable through the training of a model (Pendharkar, 2004; 2008). A learnable threshold, sensitive to the marginal costs of each error type, could not only improve classification accuracy, but also maximise the net social benefit of its outcomes (Corbett-Davies et al., 2017). Furthermore, the pupillary response data employed here does not distinguish between 'shallow' displays of acted anger and 'deep' portrayals that are grounded in real emotion for the actor (Hochschild, 1979). This research could be extended by creating a multiple class neural network classifier that is able to distinguish between shallow acting, deep acting as well as genuine displays of anger. Finally, this model does not consider the degree of anger, only its display. A neural network classifier could also be trained on ratings of veracity of emotion to consider how different levels of emotion may affect an observer's physiological response.

Bibliography

Chen, L., Gedeon, T., Hossain, MZ., Caldwell, S. (2017). “Are you really angry? Detecting emotion veracity as a proposed tool for interaction”, *OzCHI 2017.*

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A. (2017). “Algorithmic decision making and the cost of fairness.” ArXiv170108230 Cs Stat. https://doi.org/10.1145/3097983.309809

Hochschild, A. (1979). “Emotion Work, Feeling Rules and Social Structure”, *The American Journal of Sociology,* Vol. 85, No. 3, pp. 551-575.

Hossain, MZ., Gedeon, T., Sankaranarayana, R. (2016). “Observer’s Galvanic Skin Responses for Discriminating Real from Fake Smiles”, 27th *Australiasian Conference on Information Systems.*

Kogan, (1991). “Neural networks trained for classification can not be used for scoring.” *IEEE Trans. On Neural Networks.*

Kong, W., Dong, ZY., Jia, Y., Hill, DJ., (2017). “Short-Term Residential Load Forecasting based on LSTM Recurrent Neural Network”, IEEE Transactions on Smart Grid, doi: 10.1109/TSG.2017.2753802

Merity, S., Keskar, NS., Socher, R. (2017). “Regularizing and Optimizing LSTM Language Models”, arXiv:1708.02182 CS CL, https://arxiv.org/abs/1708.02182

Milne, LK., Gedeon, TD. Skidmore, AK. (1995). “Classifying Dry Sclerophyll Forest From Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum Likelihood”, *Proc. 6th Australian Conference on Neural Networks.*

Pendharkar, PC. (2004) “A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem”, *Computers and Operations Research,* Elsevier.

Pendharkar, PC. (2008). “A threshold varying bisection method for cost sensitive learning in neural networks”, *Expert systems with Applications,* Elsevier.