Long Short-Term Memory Recurrent Neural Network with Bimodal Distribution Removal for Distinguishing Anger Authenticity

Yu Rong

Research School of Computer Science,
Australian National University,
Acton ACT 2601 Australia,
u6119984@anu.edu.au

Abstract. Learning raw pupillary dilation signals from individuals can help better distinguish the authenticity of anger than verbal responses of observers. A trained machine classifier on datasets of pupillary dilation statistics has been reported to achieve high accuracy in previous researches. Considering the characteristics of time-series data, this paper builds a classification model based on the deep learning technology – Long Short-Term Memory recurrent neural networks (LSTMs), to discriminate the veracity of anger in the light of the pupillary responses from observers. Real-world experimental data is inevitable to contain noisy data points. An outlier detection technique – the Bimodal Distribution Removal (BDR) algorithm is implemented, which ideally presents improvements on performance and efficiency. However, this paper examines the effectiveness of BDR against the target anger authenticity prediction problem and shows that the BDR process has negative impacts on classification performance.

Keywords: LSTM, recurrent neural network, outlier detection, bimodal distribution removal, classification, emotion veracity

1 Introduction

Acted anger expressions attempt to convince viewers that the expressors are experiencing such angry emotions [1]. In fact, the expressors do not carry a genuine feeling. People who mistrust this pretended anger might feel sad, uneasy or even guilty, which makes the performers achieve their purpose to manipulate others' mind. The ability to discriminate the veracity of anger benefits in many situations, e.g. criminal investigation. However, earlier researches showed that the capability of human beings to consciously distinguish the veracity of anger is poor, only at 60% accuracy. The physiological signals of which people are not conscious, such as pupillary dilation, hold some information for more accurate discrimination [1]. This paper focuses on the dataset 'anger' of the pupillary dilation statistics from viewers and perform a classification task to predict whether the observed anger is genuine or posed.

In order to investigate the relationship between the input features and the target classes, machine learning techniques are widely leveraged to build the classification model, e.g. decision trees and logistic regression. However, conventional machine learning techniques only works well when the underlying assumptions on the data property and model capability are satisfied [2]. For example, it is improper to assume that the relationship between pupillary dilation statistics and the authenticity of anger is linear. To overcome this limitation existing in traditional machine learning, the neural network technology is leveraged in this classification task. Neural networks are model-free estimators, which have arbitrary decision boundaries for a classification task [3]. No particular structure or model is assumed beforehand, say, linear or quadratic decision surfaces. Neural networks are data-driven self-adaptive methods and universal functional approximators that can approximate any function and hence are more flexible in modelling real-world complex relationships [2].

Pupillary dilation is a time-varying signal where the correlation and variation trends of values are even more informatic than the values themselves. Compared to simple feed-forward neural networks, recurrent neural networks (RNNs) contain cyclic connections that make them more powerful to model sequence data and learn the contextual relationships between timestamps [4]. The dynamic unrolling mechanism makes RNNs be able to handle sequence data of variable length. In particular, the Long Short-Term Memory recurrent neural networks (LSTMs) overcome the weakness of conventional RNNs in modelling long term dependencies [5] so is widely used nowadays and is selected for this classification task.

The limitation of neural networks also exists, which can be described by a well-known problem, namely the 'bias and variance' dilemma [6]. Incorrect models lead to high bias between the desired and actual outputs, which often refers to the underfitting problem. Totally model-free inference, which overfits the training dataset, causes high variance. An ideal model should have low bias and low variance which requires plenty of data for training. Outlier removal is one of the methods to improve the balance between bias and variance when there are no sufficiently large numbers of training patterns. It reduces noise and hence variance of the model by allowing a certain degree of bias.

[7] proposed an efficient algorithm for outlier detection – Bimodal Distribution Removal (BDR). BDR investigates the behaviour of outliers by analyzing the frequency distribution of errors for all training patterns. The errors present a tendency to distribute bimodally during the training process. The low error peak contains those well-learnt patterns while the patterns falling in the high error peak are diagnosed as outliers and should be removed. BDR also provides a halting mechanism to prevent over-fitting. Real-world datasets are inevitable with noise, so is the anger dataset obtained from experiments on human beings.

This paper will investigate the effectiveness of the LSTMs in modelling time-varying pupillary dilation statistics from observers and hence classifying the anger veracity. Particularly, the efficacy of BDR technique will be examined in light of its contribution to the classification performance of the LSTM model.

The rest of this paper is as follows. In Section 2, the techniques implemented and related investigations conducted will be described. In Section 3, the experiment details and the results will be displayed along with analysis. Section 4 concludes the paper and discusses future works.

2 Method

In this section, we first introduce the anger dataset and the pre-processing operations. Then, we explain our network architecture and the optimization algorithms we used, along with the bimodal distribution removal method for outlier detection.

2.1 Dataset

In this experiment, we use the anger dataset proposed by [1] in 2017, which is obtained by actual experiments on participants of different races watching genuine or posed anger videos. Those experiments were well-designed, excluding the influence of uncertain and confounding factors, e.g. environmental light and context of video clips. Thus, the experimental results can be considered reliable. This time-series dataset contains sequences of pupillary diameters of 20 participants when watching 10 genuine videos and 10 posed videos and hence it is a balanced dataset for binary classification. Our target is to predict the anger authenticity in the video given any pupillary dilation sequence from observers. The dataset consists of three parts: (1) *PD_left*: left eye's pupillary dilation sequences for each participant (2) *PD_right*: right eye's pupillary dilation sequences for each participant (3) *Mean_PD*: average pupillary dilation sequences over all 20 participants. We choose to use *PD_left* and *PD_right* for this task.

2.2 Dataset Pre-processing

PD_left and PD_right are presented in the same form, each of which includes 20 tables corresponding to the twenty different anger videos. The table titles are the identifiers of video clips, where ids staring with the letter "T" and "F" stand for genuine and posed anger respectively. Each table consists of columns representing the pupillary dilation sequences for all participants. We deal with PD_left and PD_right using the following operations and combine the two by taking the average of them as the final pre-processed dataset for training.

2.2.1 Abnormal values correction

The original dataset has some completely empty columns caused by the absence of participants, which are meaningless and hence dropped. Also, temporary measurement errors lead to some abnormal and zero values in non-empty columns. These situations occur in random positions for a majority of non-empty columns, where a simple abandon will cause a large amount of data loss. We employ the *Tukey's method* [8] to detect abnormal values and use linear interpolate [9] to replace them. Table 1 shows a window of a typical column containing a piece of abnormal values and the results after abnormal values correction.

Tukey's method determines the susceptible abnormal data points through analyzing the values distribution. It first calculates the 1st quantile $Q1$, the 3rd quantile $Q3$ and the interquartile range IQR. Values larger than $Q3 + 1.5IQR$ or smaller than $Q1 - 1.5IQR$ will be treated as abnormal. This method is effective for the anger dataset since changes in the pupillary diameters are very small over time so the distribution of valid values is very concentrated. Sudden changes in values hence can be easily detected.

Table 1. A window of a column with a piece of abnormal values (colored in red) and the corrected version of that.

Row number	Before	After
5	16.2655	16.2655
6	15.7081	15.7081
7	16.0757	16.0757
8	9.4938	16.1137
9	0	16.1516
10	0	16.1895
11	0	16.2275
12	16.2655	16.2655
13	15.9556	15.9556
14	16.5139	16.5139

2.2.2 Normalization

Since the participants are from different races, in actual their pupillary diameters differ under normal circumstances. In order to keep the sequences consistent in scales, we adopt the min-max scaling method, for each sequence

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}},$$

where x' is the normalized value. Through this operation, sequences are transformed such that values are within a specific range [0,1]. Experimental results show that this process lifts the classification accuracy by about 6%~8%.

2.3 LSTM Architecture

The estimator for this classification task is constructed based on the Long Short-Term Memory recurrent neural networks (LSTMs). LSTMs is an improved version of conventional RNNs. An RNN can be unrolled through time to create a standard NN with the weights at each timestamp are tied together. The information transfers from previous hidden layer to the current one so the network holds memories. However, the only one hidden state vector and transition weight vector keep being overwritten when learning through long time windows. The gradients are very likely to vanish or explode so classical RNNs are very unstable and the long-term contextual information is hard to learn. LSTMs replace the simple hidden units by memory blocks that contain memory cells storing the temporal state and gates to control the information flow [5]. Such an architecture makes the network hold both long-term and short-term memories which is pretty suitable for learning the pupillary dilation sequences.

We designed a 2-layer neural network for this task, consisting of one LSTM layer and one fully-connected layer. The network has 25 input units, 32 hidden units and 2 output units. The input size should have been the number of features in the sequence, which is originally 1. However, we found that the model has difficulties in converging when fed in full-length sequences. In order to speed up training and improve model convergence, we reshape the sequences to higher dimension and shorter length. The number of hidden units is devised according to the rule of thumb and get determined through testing. The detailed testing results for different input and hidden units setting will be shown in the discussion on the hyperparameter selection in Section 3. Targeting on this binary classification problem based on the sequence data of pupillary dilations, the network can be described as a many-to-one architecture with sequence input and fixed-sized output as shown in figure 1. The fully-connected layer returning classification results is only defined at the last timestamp of the sequence.

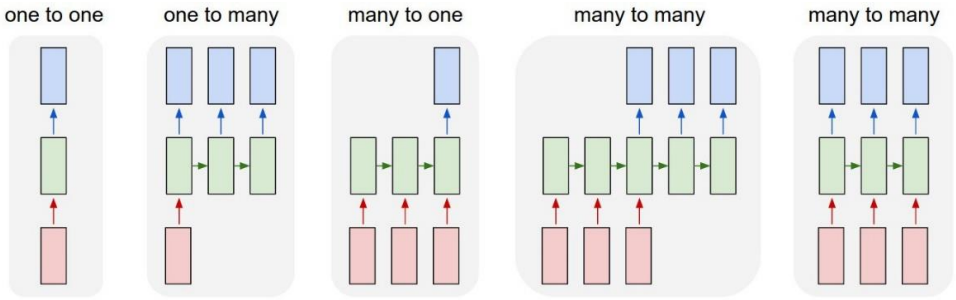


Fig. 1. Different architectures of recurrent neural networks.

2.4 Training and Evaluation

We use the optimizer 'Adam' for training. Adam is an efficient gradient-based stochastic optimizer which was confirmed to show high converging speed and require low cost of memory [10].

Furthermore, training sequences fed into the network are divided into mini-batches since mini-batch training is able to help get the algorithm out of local minima and has considerable efficiency at the same time [11]. During implementation, sequences of variable length need to be padded with zero to reach the same length if batching them. However, if the length difference between sequences is large, padding zeros even more than original sequence length will have considerable negative impacts. Feeding pack padded sequence instead of only padded tensor into the network becomes a solution to handle variable length batch training. Such a data structure packs the zero-padded batch sequences and the list representing the valid length of the sequences together, and feeds it into the network. As a result, the output is derived from the last valid timestamp, which eliminates the bad effects caused by long zero paddings.

We evaluate the cross-entropy loss in the back-propagation process. Cross-entropy loss penalizing those confident but wrong predictions especially, is commonly used for classification tasks [12]. For binary classification, the cross-entropy loss is defined as

$$Loss = -(ylog(p) + (1-y)\log(1-p)),$$

where y is the binary indicator 0 or 1 of the desired label and p is the predicted probability of the desired class.

2.5 Bimodal Distribution Removal

Bimodal Distribution Removal is an outlier detection method proposed by P. Slade and T.D. Gedeon [7]. It confirmed that networks have the ability to identify outlier patterns itself by analyzing the error distribution among all training patterns. For most of the training processes, the pattern errors tend to have an approximately bimodal distribution after a short term of training. The low error peak contains the well-learnt patterns while the network feels strenuous to learn the patterns appearing in the high error peak. Under this condition, some of those high error patterns are diagnosed as outliers and need to be abandoned. Analysis on the variance of pattern errors is employed to identify the formation of the bimodal distribution. Low variance v (≤ 0.1) indicates that a majority of patterns have been learnt well and hence the bimodal distribution takes shape. The detailed removal process can be explained in two steps.

(1) Screen outlier candidates: The mean pattern error $\bar{\delta}$ that got magnified by the high error mode, will be larger than nearly all patterns in the low error mode. Select patterns with error $\geq \bar{\delta}$ as a subset of outlier candidates, which actually filters out patterns in the low error mode.

(2) Isolate potential outliers: The error distribution of the candidate subset will skew to the outliers. Calculate the error mean $\bar{\delta}_{ss}$ and the error variance v_{ss} of the subset and remove all patterns with

$$\text{error} \geq \bar{\delta}_{ss} + \alpha \cdot v_{ss}\ , \text{ where } \alpha \in [0,1]$$

from the training set.

The BDR process repeats every 50 epochs and halts training when a small enough variance $v \leq v_h$ (0.01 typically) occurs, indicating the well-trained network. This halting mechanism prevents the network from over-fitting the gradually smaller training set and improves training efficiency.

Networks tend to have different converging speed due to distinct constructions. Unlike the 200~500 epoch reported in [5], our network usually formats the desired bimodal distribution at epoch 8~12. Figure 2 presents the pattern error distributions at epoch 9 and 11. It can be observed from the figures that a bimodal distribution appears with the low error peak at loss ≈ 0.0 and the high error peak at loss ≈ 1.5.

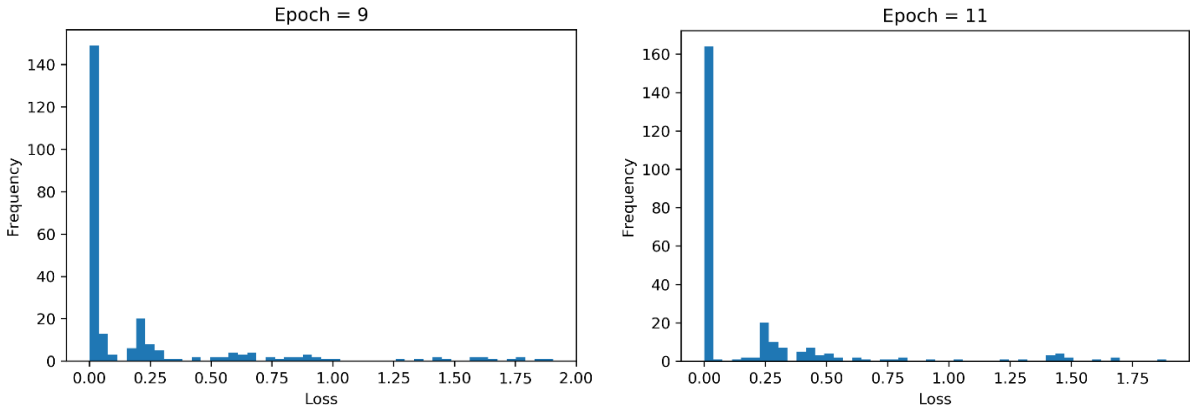


Fig. 2. Histograms for the pattern error distributions at epoch = 9 and 11.

We add another complementation to adapt the BDR algorithm in this binary classification problem. It is insufficient to determine the time when the BDR process starts simply by monitoring variance changes. An unexpected finding was that the initial variance of pattern errors is below 0.1 at the first 0~5 epochs. The possible reason is that any plain network, with zero knowledge base, is expected to give the right prediction with a 50% probability for a binary classification problem. Hence the error distribution tends to be concentrated at the beginning of training. To avoid shrinking the training set from the very beginning, the first 5 epochs are ruled out from the BDR process when realizing the algorithm.

3 Results and Discussion

In this section, we explain the experiments in details, including the tests for hyperparameter selection. We present the experimental results of our neural network model trained with and without bimodal distribution removal and compare them with the results published in research papers on the anger dataset.

3.1 Experimental Setup

We compared the performance of our model with and without the BDR technique against the results published in a research paper [1] on the anger dataset. The construction of the network and the implementation of the training process and algorithms are carried out using Pytorch framework [13] under Python. We use 10-fold cross validation to test the classification performance. In order to ensure a fair comparison and exclude random effects of the fold assignments and batch partitions, the experiments are repeated 10 times and identical random seeds are set for each experiment in one control group. Average prediction accuracies with their standard deviations are reported.

3.2 Hyperparameters

Since the sequences are reshaped to have the dimension equal to the input size, larger input size leads to shorter length of sequences and hence stronger convergence ability of the model. But we still need to ensure a certain sequence length to learn the contextual information; otherwise the LSTM layer makes no sense. The number of hidden units determining the upper bound of the model complexity, impacts the learning ability of the network significantly. A proper setting for the hidden size is necessary to prevent the under-fitting and over-fitting problems and to ensure better generalization ability of the model. According to the rules of thumb, the hidden size ought to be less than the total number of training samples. Through experiments, small changes in the number, e.g. one or two, give rise to a negligible difference in test accuracy. Table 2 provides the testing results on some typical numbers of input and hidden units, where the number of epochs is independently tuned to preserve best test accuracy. What stands out in this table is the relatively highest and fairly stable test accuracy when there are 25 input neurons and 32 hidden neurons.

Table 2. Average test accuracy (± standard deviation) over five repeated experiments of random train-set-and-test-set split and random partition of batches, when the learning rate equals 0.1 and the batch size is 64.

Input size	Hidden size	Accuracy (%)
15	**20**	88.12(±2.50)
20	25	92.88(±1.56)
20	32	93.10(±1.72)
25	25	93.60(±2.05)
25	**32**	**94.36(±1.42)**
30	32	93.87(±1.05)

The learning rate is set at 0.1 and the batch size is 64. Our network with 25 input units and 32 hidden units turns out to perform best predictions when training iterates for about 300 epochs. At that timestamp, the model fully converges and is still distant from overfitting.

We also carry out some tests to select the best hyperparameter α for the BDR process. Table 3 provides the average test accuracy under different α among ten test runs, showing that the best performance occurs when $\alpha = 0.8$.

Table 3. Average test accuracy (± standard deviation) over 10 repeated test runs for different α.

α	Accuracy (%)
0.6	86.49(±2.79)
0.7	88.74(±3.03)
0.8	**90.22(±3.35)**
0.9	89.11(±2.86)
1	88.54(±3.42)

3.3 Results

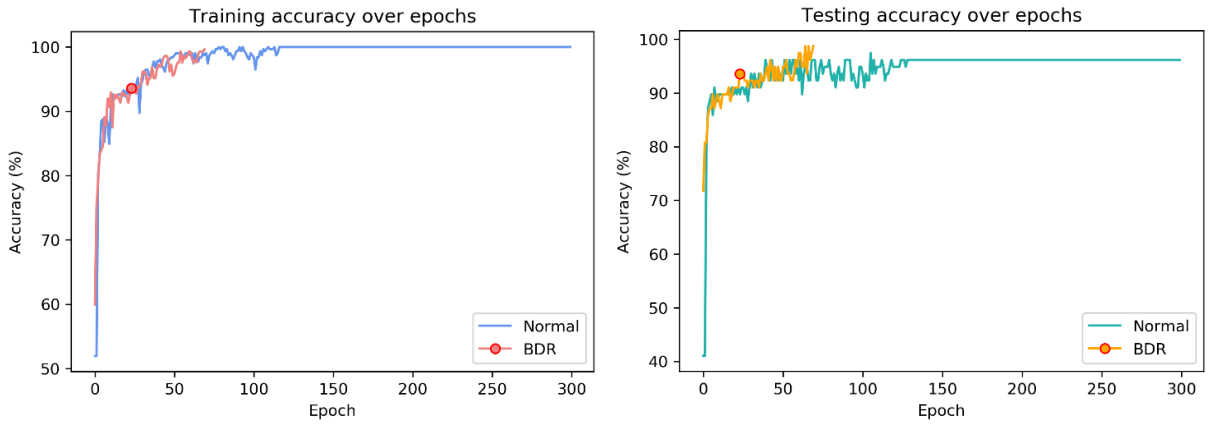
In this section, we report the results of our LSTM model on the anger dataset with and without the BDR process and compare classification accuracy of our model against the results in the dataset paper [1] and the Misaka neural network model proposed by [14].

LSTM with and without BDR. We have two variants of our LSTM model, training with BDR and without BDR. With the assistance of BDR, the model can occasionally finish training early at 80~150 epochs. Figure 3 displays three different situations that occurred in experiments, where the BDR has positive, negative or nearly indifferent impacts on the final performance respectively. The negative situation appears more frequently than the positive one and the indifferent one rarely occurs during a large number of experiments. Table 4 provides the average test accuracy for neural network models trained with and without BDR among ten test runs. The normal LSTM is about 4% more accurate than the LSTM trained with BDR and BDR training presents more severe instability. Both of them have higher accuracy than the verbal responses (60%).

Table 4. Average test accuracy (± standard deviation) over 10 repeated test runs for NN with and without BDR.

Model	Accuracy (%)
LSTM	94.87(±1.66)
LSTM with BDR	90.56(±3.86)

(1)



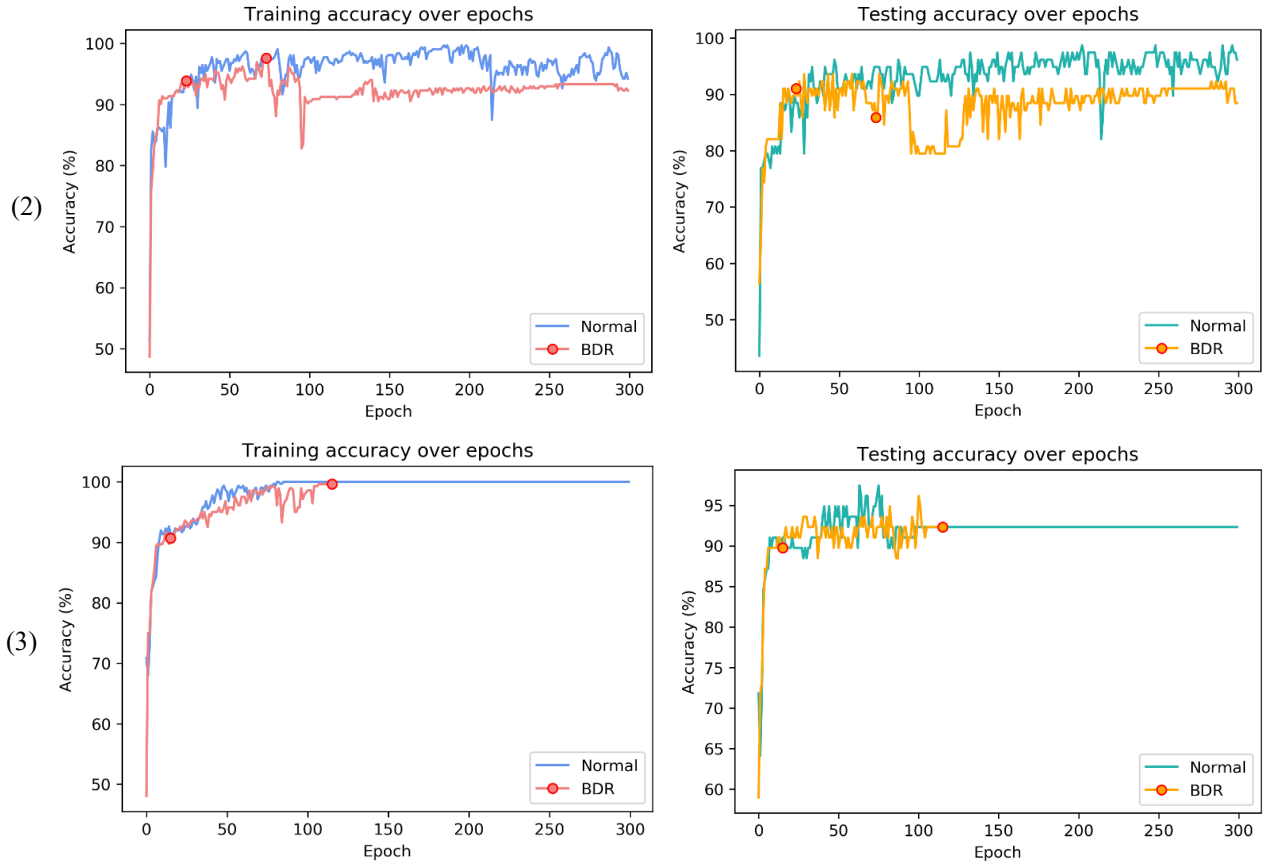


Fig. 3. Figure groups (1), (2), (3) represent the positive, negative and indifferent situations respectively. Line charts on the left column show the training accuracy tendency over epochs. Line charts on the right column show the testing accuracy tendency over epochs. Red circles on BDR lines stand for the timestamps when BDR is triggered.

Our model against published results. The anger dataset paper reports a classification accuracy of up to 95% but without explaining the concrete implementation methods. So, we refer to another Misaka neural network proposed by Qin and Gedeon et. al. that has achieved good prediction results for the anger dataset in recent years. Table 5 provides an intuitive comparison between the average test accuracy of our model, the Misaka neural network and the dataset paper. It is obvious that our LSTM model significantly outperforms the Misaka neural network and almost reaches the accuracy reported in the anger dataset paper.

Table 5. Average test accuracy (± standard deviation) of our LSTM model and the Misaka NN, and the result reported in the anger dataset paper.

Model	Accuracy (%)
Our LSTM	94.87(±1.66)
Misaka NN	88.9
Dataset Paper	95

3.5 Discussion

Although it is undeniable that BDR sometimes accelerates training and prevents over-fitting, BDR process presents no significant improvements in the overall prediction accuracy and have very unstable effects. Without doubt, the LSTM with BDR is inferior to the normal LSTM by about 4% of average accuracy.

In fact, the judging mechanism based on the error variance is insufficient in handling various situations. For this binary classification problem, the error distributions usually tend to be simply right-skewed after certain epochs and won't come out with obvious double modes at most of the time but BDR cannot distinguish them. When the bimodal distribution is not evident, the errors of the outlier candidates isn't left-skewed, under which a large part of the data points that are tragically abandoned won't be outliers. Then the network will lose some features to learn and have poor prediction ability on unseen datasets. Moreover, the halting timestamp determined by a predefined threshold of variance cannot guarantee that the model jumps out of training in its best status and sometimes causes training to end prematurely.

4 Conclusion

This paper constructs an LSTM recurrent neural network model to predict the authenticity of expressors' anger, through learning the pupillary dilation sequences obtained from observers. This paper verifies that learning unconscious physiological signals from observers through neural networks can achieve high accuracy in distinguishing the veracity of anger, compared to the verbal responses. It also compares the model we devised with the Misaka neural network and the results reported in the dataset paper. It declares that our model transcends the Misaka NN and almost reproduces the results reported in the dataset paper. LSTMs' capability in learning sequential data and contextual features is fully confirmed. This paper also examines the effectiveness of the bimodal-distribution-removal outlier detection algorithm in this binary classification problem. We conclude that the BDR process actually has no significant improvements in the prediction performance on the anger dataset and sometimes negatively affects training.

Potential future works can focus on the following two points. First, we can improve the network architecture to make it own a stronger ability to handle longer sequences, e.g. deep LSTMs, so that long sequences will not need to be reshaped into multiple dimensions and hence more contextual information are preserved. Second, targeting at the defects of BDR appearing in this task, we can devise some extensions to it: (1) a more intelligent method to distinguish the bimodal distribution (2) a more flexible halting mechanism.

References

[1] Chen, L., Gedeon, T., Hossain, M., Caldwell, S.: Are you really angry? Proceedings of the 29th Australian Conference on Computer-Human Interaction - OZCHI '17. (2017).

[2] Zhang, G.: Neural networks for classification: a survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews). 30, 451-462 (2000).

[3] Hepner, G., Logan, T., Ritter, N. and Bryant, N..: Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification. Photogrammetric Engineering and Remote Sensing, 56(4), pp. 469-473. (1990)

[4] Sak, H., Senior, A. W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. (2014)

[5] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation. vol. 9, no. 8, pp. 1735–1780. (1997)

[6] Geman, S., Bienenstock, E., Doursat, R.: Neural Networks and the Bias/Variance Dilemma. Neural Computation. vol. 4, pp. 1-58 (1992).

[7] Slade, P., Gedeon, T.D.: Bimodal Distribution Removal. International Workshop on Artificial Neural Networks. pp. 249-254. Springer (1993)

[8] Seo, S.: A review and comparison of methods for detecting outliers in univariate data sets. Doctoral dissertation, University of Pittsburgh. (2006)

[9] Blu, T., Thevenaz, P., Unser, M.: Linear Interpolation Revitalized. IEEE Transactions on Image Processing. vol. 13, pp. 710-719 (2004).

[10] Kingma, Diederik, Ba, Jimmy.: Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, San Diego. (2015)

[11] Li, M., Zhang, T., Chen, Y., Smola, A.: Efficient mini-batch training for stochastic optimization. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14. (2014).

[12] Janocha, K., Czarnecki, W. M..: On loss functions for deep neural networks in classification. Theoretical Foundations of Machine Learning. (2017)

[13] Paszke, A., Gross, S., Massa, F. et. al: PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems. vol. 32. pp. 8024—8035. (2019)

[14] Qin, Z., Gedeon, T., Chen, L., Zhu, X., Hossain, M.: Artificial Neural Networks Can Distinguish Genuine and Acted Anger by Synthesizing Pupillary Dilation Signals from Different Participants. Neural Information Processing. pp. 299-310 (2018).