# Feature selection models for detecting depression levels based on Brute force analysis

Tianhao Zhao

*Research School of Computer Science*
*The Australian National University*
Canberra, Australia
u6815967@anu.edu.au

**Abstract.** Depression patients are considered to behave different from normal people. This experiment utilises neural networks to detect the relationship between depression levels and physiological signals. Besides it also implement a Brute force approach to analyse the significance of input features for the neural network. The results show that feature selection based on Brute force analysis helps the neural network improve about 5%.

**Keywords:** Brute Force, Feature Selection, Depression Detection, Physiological Signals, LSTM.

## 1 Introduction

Human beings are emotional creatures. People laugh, cry or shout driven by different feelings such as happiness, sadness, anger or fright. These emotions are essential components of our mental world, providing us with a colorful life. However, not all emotions are beneficial. Some bad moods which deteriorate our life quality are not welcomed.

Depression is among these unwelcome emotions. Depression is a common mental disorder together brings people loss of interest and enjoyment, low self-worth, poor concentration and appetite [1]. Depression can affect people's daily routines and even leads to suicide. It is predicted by WHO that depression would become the second largest global disease in 2020 [2].

Therefore, detecting depression is essential. Scientists have been working on this topic for long. Recent advances utilises physiological signals to help diagnose depression [3]. Patients with depression, are found to behave differently from other people, especially in physiological signals including galvanic skin response [4], pupillary dilation [5] and skin temperature [6].

In order to assist diagnose depression, Neural Networks are used to detect the potential relationships between the physiological signals and depression level. The input features are obtained from Zhu's dataset [7]. Besides, Brute Force algorithm [8] is also implemented to analyse the importance of different features extracted from the dataset.

This paper introduces the physiological signals used in the Neural Networks and Brute Force algorithm. In experiment part it reports the results of different physiological signals and the importance of each feature to the network. It improves the network with feature selection based on brute force. It also implemented a LSTM [9] model training on the timestamp recordings of phychological signals. The paper concludes the findings in the experiments and proposes future expectations.

## 2 Methodology

### 2.1 Preprocessing

The dataset includes three physiological signals, Galvanic Skin Response (GSR), Pupillary Dilation (PD) and Skin Temperature (ST). They are preprocessed in advance by Zhu [7]. The raw data is a timeseries recording of three physiological signals of participants. For PD only there is an interpolation process due to miss of data caused by occasional blinks. After normalisation, the signals are smoothed and segmented. A total number of 85 features are extracted and saved in three excel files, 23 features for GSR and ST respectively and left 39 features for PD.

The extended version of raw timeseries records is also provided. Just as Zhu's prepocessing, the raw data is normalised and interpolated. We extract each video from one participant as a row data and normalise it. Replace all 'Nan' type with 0 and concatenate altogether with the label information, we will get a similar dataset with the former one. This timeseries data is used in a LSTM model, which will be later introduced.

### 2.2 Neural Network

This paper utilises a simple fully connected neural network to train the extracted features for classifying depression levels. The network has a 50 hidden-neuron layer. The input is extracted features from different physiological signals and the output is number representing the depression level 'None', 'Mild', 'Moderate' and 'Severe'. The network uses Cross-Entropy loss and Adam optimiser.

To evaluate the performance of the trained model, besides basic loss and train accuracy curves and validation accuracy, the experiment also chooses precision, recall and f1-score as measures. Since depression has four levels, all measures are calculated on each level respectively.

During the experiment, it is found that the measures fluctuate heavily. Considering there are only 192 samples in the dataset and test samples used for measures is fewer, the fluctuation is explainable. Therefore, we run for 10 times on each signals and compute the average for all measures.

### 2.3 Brute Force

One way to determine which features are the most important in the trained model is Brute force approach [8]. Brute force approach eliminates two features from the original input, and calculate the total sum of squares (tss) value of prediction on test set. Each run of eliminating a pair of features derive a tss value, so that if there are $n$ features, there are $\frac{n(n-1)}{2}$ ways of eliminating pairs of features, thus the same number of tss values.

Ranging these values in increasing order and plot them, we will get a discontinuous curve with some gaps. The greater the gap is, the more significant the eliminated pair of features is. The longer a continuous stable line is, the less significant the eliminated pair of feature is. Therefore, we can determine the order of significance of features through the brute force graph.

### 2.4 Features selection

Predictions on depression level are based on extracted features from the physiological signals. However, those extracted features may contain something redundant or irrelevant. The significant features would do less contribution to the output, therefore affect the final performance. Hence, feature selection is necessary.

Since brute force can derive the rank of the inputs' significance, the network can be modified and improved with this information. It is a good approach to remove redundant or irrelevant features from the network and improve its efficiency. So the next step is invalidating such unnecessary features based on what brute force finds. The experimental performance will be covered later in this paper.

### 2.5 LSTM

For an extended version, we are given the raw timestamp data for all participants. Therefore, the temporal information can be utilised with a LSTM [9] model. The experimental model contains 3 lstm modules. In order to avoid overfitting, each lstm module follows by a dropout layer. Instead of extracted features as inputs in neural network, this time we use the preprocessed temporal data as inputs directly.

The basic settings for this model is same as the Neural Network, which includes 50 hidden neurons for each module, Cross-Entropy loss and Adam optimiser.

# 3 Result

## 3.1 Neural Network

The experiments are based on four conditions, comprised of three physiological signals and the combined one. Table 1 shows the measures of precision, recall and f1-score for different depression levels in three physiological signals respectively. The average precision are 0.43, 0.35 and 0.40 for GSR, PD and ST features. The average recall are 0.47, 0.34, 0.41 and average f1 score are 0.45, 0.35, 0.40 respectively.

As for four depression levels, it seems that in GSR signal, the 'Severe' level is easier to detect compared with the medium two levels. Other two signals don't have such phenomenon.

Figure 1 shows the average test accuracy of each physiological signals and the combination on 10 runs. We can see that GSR performs best at 0.48 while PD performs worst at 0.34. The accuracy of ST is 0.39 while the combination reaches 0.42. Although the performance does not reach a high level, it is better than random guessing (0.25).

**Table 1.** Performance measures for each depression level on all features.

| Physiological signal | Galvanic skin response | | | Pupillary dilation | | | Skin temperature | | |
|---|---|---|---|---|---|---|---|---|---|
| Depression Level | Precision | Recall | F1-score | Precision | Recall | F1-score | Precision | Recall | F1-score |
| **None** | 0.45 | 0.52 | 0.48 | 0.36 | 0.34 | 0.35 | 0.40 | 0.40 | 0.40 |
| **Mild** | 0.38 | 0.34 | 0.36 | 0.30 | 0.34 | 0.32 | 0.42 | 0.38 | 0.40 |
| **Moderate** | 0.33 | 0.40 | 0.37 | 0.39 | 0.37 | 0.38 | 0.41 | 0.45 | 0.43 |
| **Severe** | 0.57 | 0.63 | 0.60 | 0.35 | 0.32 | 0.33 | 0.37 | 0.39 | 0.38 |
| Average | 0.43 | 0.47 | 0.45 | 0.35 | 0.34 | 0.35 | 0.40 | 0.41 | 0.40 |

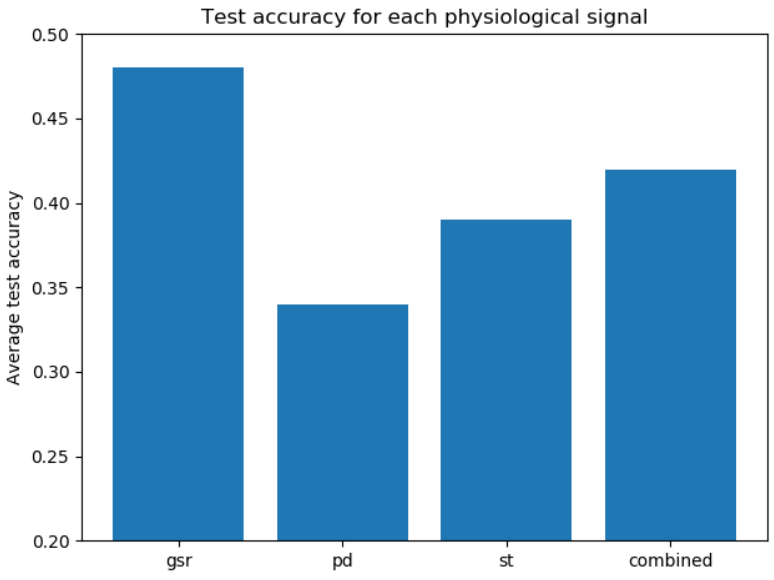
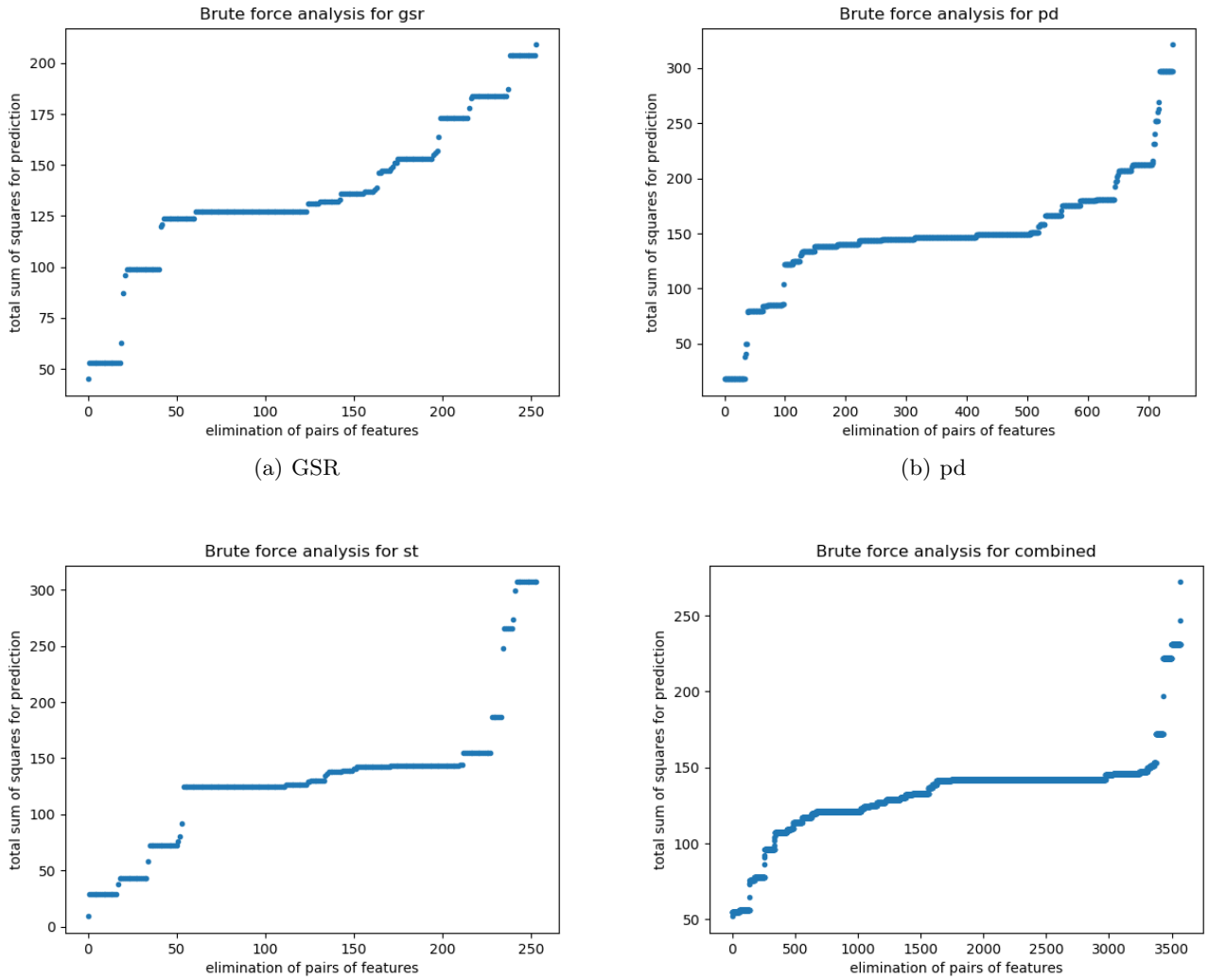



**Fig. 1.** Test accuracies for each physiological signal

### 3.2 Brute force analysis

Figure 2 shows the results on tss values for different physiological signals. Notice that GSR and ST signal contain 23 features and PD signal contains 39 features, so that the total number of pairs of eliminated features are different. As we can see in the figure, there does exist some gaps in these physiological signals, which means that certain inputs do contributes more to the final output while others contributes less.




(a) GSR (b) pd

(c) ST (d) combined

**Fig. 2.** Brute force analysis for input features for four conditions of physiological signals

In order to explore which inputs are most significant and which are least significant, the tss values of all pairs of eliminated inputs is recorded in a square matrix with dimension of $n \times n$, where $n$ denotes the number of inputs. The $ij$-th element of the matrix denotes the tss value of eliminating the $i$-th and $j$-th inputs. Then the tss values are summed up in rows, so that a total tss value for one input is acquired. Ranking them in descending orders, the significance of each input is indirectly obtained. Table 2 shows the result of one try.

Considering that the training progress of neural networks is accompanied with randomness, more attempts are picked and the changes of 5 most significant inputs for each physiological signals are shown in Fig 3.

**Table 2.** Significance rank of features in each physiological signal

| Physiological signal | Most significant | | | | | ··· | Least significant | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GSR | 16 | 18 | 0 | 22 | 13 | ··· | 3 | 20 | 17 | 1 | 19 |
| PD | 27 | 22 | 29 | 5 | 34 | ··· | 36 | 21 | 10 | 9 | 35 |
| ST | 6 | 11 | 8 | 7 | 20 | ··· | 10 | 0 | 16 | 19 | 17 |

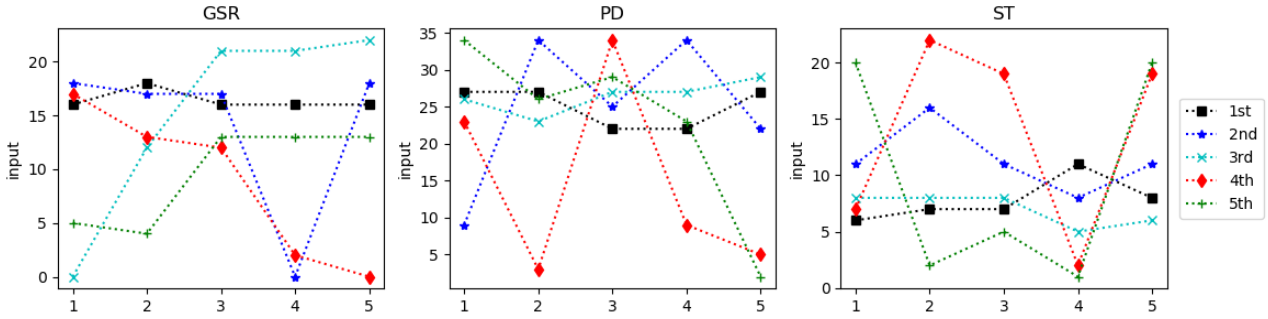



**Fig. 3.** Changes in the 5 most significangt inputs on overtraining on each physiological signal

Generally, some inputs appear continuously in all attempts regardless of the order, suggesting that they do influence the outputs significantly. For example in PD signal, the 22-th, 27-th and 34-th input appear frequently, showing their dominant positions against other inputs. However, the randomness of the training progress is still unavoidable. The significance of certain inputs are still not robust to predict in one training progress. Take ST as an example, the 19-th input is regarded as one of the least significant inputs in Table 2, while in Fig 3, it appears twice within top 5 significant inputs. That just indicates that in one training, the significance of one input is not determinate. Although there might be some potential preference, it is not guaranteed that each time it will emphasize on certain inputs.

### 3.3 Features selection

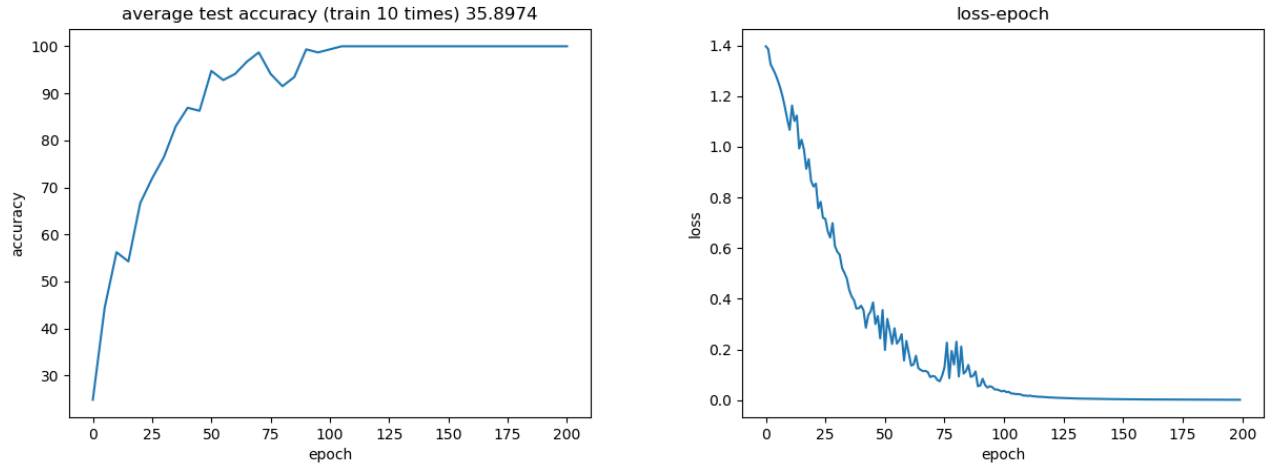
Based on the significance brute force analysed before, some less significant inputs can be dropped while those more significant inputs are left. Therefore the network could be more effective. Take GSR dataset as an example. GSR contains 23 input features. Applying brute force analysis the rank of those input features' contribution is derived. Experiments show that the best number of enabled inputs for GSR is around 5. The final test accuracy of the model raises to 0.53. The detailed measures of this experiment compared with those of original network is shown in Table 3

**Table 3.** Performance measures for original network and brute force network on GSR, with 5 features enabled.

| Algorithms | Neural Network | | | Brute Force Network | | |
|---|---|---|---|---|---|---|
| Depression Level | Precision | Recall | F1-score | Precision | Recall | F1-score |
| **None** | 0.45 | 0.52 | 0.48 | 0.49 | 0.55 | 0.52 |
| **Mild** | 0.38 | 0.34 | 0.36 | 0.41 | 0.52 | 0.46 |
| **Moderate** | 0.33 | 0.40 | 0.37 | 0.46 | 0.35 | 0.39 |
| **Severe** | 0.57 | 0.63 | 0.60 | 0.65 | 0.60 | 0.62 |
| Average | 0.43 | 0.47 | 0.45 | 0.50 | 0.50 | 0.50 |

### 3.4 LSTM

The performance of LSTM model is not so good. Take GSR feature as an example, the average test accuracy is 0.359 on 10 runs. Although dropout layers have been added, it seems that the influence of overfitting cannot be eliminated. The training accuracy raises quickly to top within 100 epoch, while the test accuracy still fluctuates between 0.25 to 0.5.




(a) accuracy (b) loss

**Fig. 4.** accuracy and loss curves of LSTM model

## 4 Discussion

The above experiments includes simple neural network, brute force analysis, feature selection and LSTM model. The former three are based on dataset of extracted features while the last one is based on normalised timesteps dataset. The best performance of simple neural network is 0.48 on GSR dataset. The brute force analysis indicates that among those extracted features in three phychological signals, some features are redundant or irrelevant. Based on brute force analysis, the neural network improves with about 0.05 after feature selection, reaching 0.53 on GSR dataset. The LSTM model fails to overtake the performance of neural network, only reaches 0.36 on GSR dataset. Compared with state of art methodologies, there is still a long way to go. In regard to the poor performance, the limitation of samples in this dataset should be to blame. Although it provides 85 features, the total sample number is only 192, despite that it should be used for both training and validating. However, such kind of dataset can be hardly augmented. Traditional transformer cannot be used in this dataset as features are extracted from records of phychological signals. If simply adding small bias on these samples to acquire extra data, the robustness of the model is doubting. Generating similar samples with GAN has the same worry.

The Brute force approach helps indicate how significant the input features are to the results. This information helps us to do feature selection. To some extent, it achieves a similar function as Genetic Algorithm. One limitation for brute force is that it only reflects the rank of input features' significance indirectly through tss values. It could not reflect the exact contributions those features do to the network. Hence, there is no clear thresholds for whether one features should be enabled or not. The enabling process needs to be manually set.

In dealing with raw timeseries data, a LSTM model is implemented. However, it fails to exceed the performance of nerual network. The curves show that the model may fail into overfitting. It might concentrate more on rigid trends of phychological signals over timestamps in training set and fails to adjust to more general case.

## 5 Conclusion

In this experiment a neural network model is implemented for detecting depression levels from physiological signals. The best accuracy is 48% on GSR dataset. After doing feature selection based on brute force approach, the best accuracy rise up to 53%. For attempts on timestamp dataset, the LSTM model performs at 36%.

For further improvement, one thing is data augmentation. With more samples, the performance would be better. As for LSTM model, how to avoid overfitting needs further exploration.

## References

1. Marcus, M., Yasamy, M. T., van Ommeren, M. V., Chisholm, D., Saxena, S. (2012). Depression: A global public health concern.
2. Murray, C. J., Lopez, A. D., World Health Organization. (1996). The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary. World Health Organization.
3. Chen, Y. T., Hung, I. C., Huang, M. W., Hou, C. J., Cheng, K. S. (2011, October). Physiological signal analysis for patients with depression. In 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI) (Vol. 2, pp. 805-808). IEEE.
4. Greenfield, N. S., Katz, D., Alexander, A. A., Roessler, R. (1963). The relationship between physiological and psychological responsivity: depression and galvanic skin response. Journal of Nervous and Mental Disease.
5. Siegle, G. J., Steinhauer, S. R., Thase, M. E. (2004). Pupillary assessment and computational modeling of the Stroop task in depression. International Journal of Psychophysiology, 52(1), 63-76.
6. Lin, H. P., Lin, H. Y., Lin, W. L., Huang, A. C. W. (2011). Effects of stress, depression, and their interaction on heart rate, skin conductance, finger temperature, and respiratory rate: sympathetic-parasympathetic hypothesis of stress and depression. Journal of clinical psychology, 67(10), 1080-1091.
7. Zhu, X., Gedeon, T., Caldwell, S., Jones, R. (2019). Detecting emotional reactions to videos of depression. In INES'19: IEEE 23rd International Conference on Intelligent Engineering Systems (6 pp).
8. Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. International Journal of Neural Systems, 8(02), 209-218.
9. Gers, F. A., Schmidhuber, J., Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.