Genetic Algorithm for Feature Selection - Detecting Emotional Reactions to Videos of Depression

Sin-Jie, Wu
Research School of Computer Science, Australian National University
u6884288@anu.edu.au

Abstract: Using the collected physiological responses data from 12 experiment participants who watched video recordings of 16 individuals suffering from various levels depression to examine if observers' physiological responses could reflect other individuals' depression level. From the results to build a neural networks classifier for identifying depression levels based on features derived from observers' physiological responses, which are Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD). The built NN and Casper NN models are applied Genetic Algorithm for feature selection to improve model accuracy and efficiency. Finally, a comparison of the models is performed based on the accuracy, precision and recall in order to identify the best classification method that allows the detection of depression level based on physiological responses. According to the results, the best performance is presented by the Casper NN model with Genetic Algorithm for feature selection, obtaining an accuracy of 62.5%, which allows to conclude that it is possible to automatically classify depression level based on a reduced set of features with significant accuracy.

Keywords: Depression Detection, Physiological Signals, Galvanic Skin Response, Skin Temperature, Pupil Dilation, Casper Neural Network, Genetic Algorithm, Feature Selection

1 INTRODUCTION

Nowadays, more and more people suffer from depression, stress and anxiety. They may seek the help from doctors. However, current diagnosis almost relies on self-reported questionnaires or clinician assisted interview which are sometimes too subject to make a prediction. Therefore, the main idea of the experiment is to detect and determine the level of the depression from video recordings of participants expressing themselves with labelled depression level. The goal is to know whether a classifier could be developed to recognize other individuals' depression level using observers' physiological signals, including Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD). Referring to the paper "Detecting emotional reactions to videos of depression" (X. Zhu, T. Gedeon, S. Caldwell and R. Jones, 2019) and "Casper Neural Network Implementation - Detecting Emotional Reactions to Videos of Depression" (S. Wu, 2021), the researchers built neural networks based classification models with Adan optimizer by using backpropagation with the Cross-Entropy loss function and also finetuned the original model with Casper algorithm. In addition, taking advantage of precision, recall and F1-score for model evaluations. In this paper, we focus on reproducing the experiment using the collected pre-processed data from the original experiment and making use of genetic algorithm for feature selection to examine if the testing accuracy and evaluation are improved or not with less but more representative subset of features.

We used Neural Networks and Casper Neural Network to recognize depression levels based on features derived from observers' physiological responses to others' depression. Actually, the classification performance can be improved further by selecting the most influential features during the training phase. Furthermore, not all of these features are equally important for a classification decision. Hence, there is still a need for more effective feature selection methods, which can contribute towards improving the performance of depression level detection. This encourages us to apply evolutionary algorithm (EA) based feature selection in order to enhance the depression level detection. We applyied genetic algorithm (GA) to select subsets of features for optimizing depression classifications since GA has been successfully applied to select features from physiological signals (N. Sharma and T. Gedeon, 2013). Furthermore, considering the huge number of features compared to the modest number of samples in these datasets, other researchers have shown that GA resulted in significantly better classification accuracy (S. Sayed, M. Nassef, A. Badr, I. Farag, 2019). Approaches for depression recognition of video watchers are developed and discussed, including a method for selecting optimally useful features from the response signals. The paper concludes with a summary of the findings and suggests directions for future work.

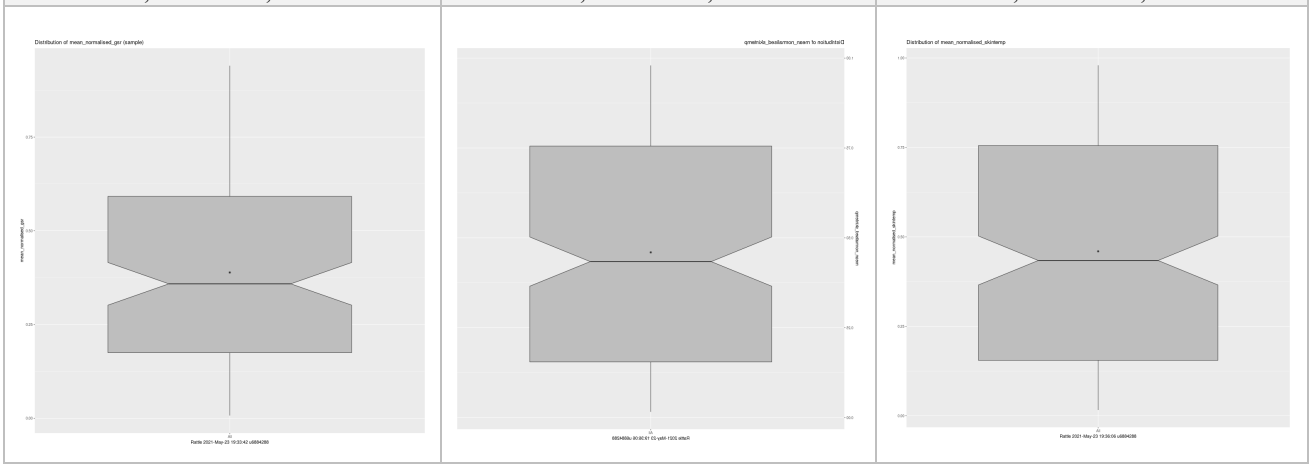
2 METHODOLOGY

A. Data Inspection and Data Preparation

The collected data contains 200 complete responses from participants including their conscious depression judgements (none, mild, moderate and severe) and physiological sensors recordings. We used the previously collected physiological signals to help diagnose individuals' depression level. There is total 14 participants participated in the experiment, however, only 12 participants' result are recorded successfully at the end, therefore only 192 records in every dataset. Total three separate datasets which are Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary

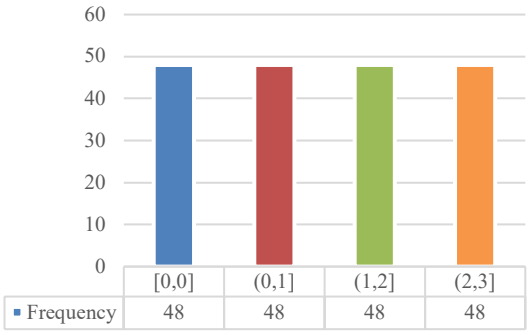
Dilation (PD) were used in the training model procedure. Without raw data, taking pre-processed normalized data is a better mean to make sure all datasets take on similar range of values so that gradient descents can converge more quickly. From statistical aspect, we compared the mean normalized features across each segmented physiological data set and the result is show by box plots (Figure 1). It is clear that the dispersion of data distribution is similar between ST and PD; while GSR is more concentrated distributed that is why we used the normalized data to deal with this issue in raw data.

Fig. 1. Box Plots of Mean Normalised Features

Mean Normalised GSR	Mean Normalised ST	Mean Normalised PD
Min; Median/Mean; Max 0.01; 0.36/0.40; 0.94	Min; Median/Mean; Max 0.02; 0.43/0.46; 0.98	Min; Median/Mean; Max 0.18; 0.38/0.40; 0.75
		

Furthermore, we know that the distribution of the classes of the depression label is fairly well balances, every class has 48 responses (Figure 2), showing that there is no bias issue in our training datasets. The four depression categories specifying that: 0 (none): indicates no or minimal depression; 1 (mild): indicates mild depression; 2 (moderate): indicates moderate depression; 3 (severe): indicates severe depression.

Fig. 2. Visualize Distribution of Depression Label in Every Dataset



B. Features Extraction

There is total 23 features extracted from normalized and filtered GSR signal, 23 features from ST, and 39 features from PD, mainly focusing on capturing the amplitude variance and the occurrences of transient changes in the signals. The minimum, maximum, mean, standard deviation, variance, root mean square, means of the absolute values of the first and second difference were calculated as features. Therefore, there are 85 features from the three physiological signals: 23 (GSR) + 39 (PD) + 23 (ST) in total.

C. Features Selection

It is not efficient to train a neural network classifier with a large amount of features due to expensive computation cost. In addition, features selection also enhances generalization by reducing overfitting and avoids the curse of dimensionality. Since our data contains some features that are either redundant or irrelevant and can thus be removed

without incurring much loss of information. However, feature selection through brute force trial and error would require a lot of computation cost; thus, we tried to improve upon the method by applying a genetic algorithm to evolve a network with the goal of achieving optimal hyperparameters in a fraction the time of a brute force search. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection. Using presence (1) or absence (0) of every possible feature in the model (Figure 3), based on the linear regression performance of a classifier as the fitness function due to numerical datatype in all features (Tong, D.L., Mintram, R., 2010). The initial population for the GA was all 85 features; then setting a chromosome with binary string format to represent a feature was used for classification or not. All settings for GA used in the hybrid classification system are shown in Table 1. Each gene (bit) is selected randomly from one of the corresponding genes of the parent chromosomes, so uniform crossover and uniform mutation are chosen to be used in this experiment. Furthermore, tournament selection was used for selecting the fittest candidates (chromosomes) from the current generation in a Genetic Algorithm. These selected candidates were then passed on to the next generation for crossover.

Fig. 3. GA representation of features.

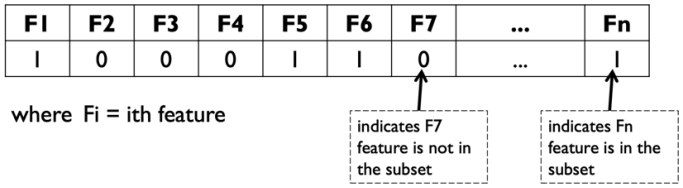


Table 1. Implementation Settings for GA

GA Parameter	Value
Population size	100
Generation size	100
Mutation probability	0.1
Crossover type	Uniform crossover
Mutation type	Uniform mutation
Selection type	Tournament Selection
Tournament size	2

D. Neural Networks Based Classification Models – NN vs. Casper NN (with GA for each model)

In this paper, we built four Neural Network (NN) based depression classification models:

I. NN: a NN classification model that used all depression features as input to recognize depression patterns.
II. GA+NN: a NN that used a subset of features selected by a GA to recognize depression patterns.
III. Casper NN: a Casper NN classification model that used all depression features as input to recognize depression patterns.
IV. GA+Casper NN: a Casper NN that used a subset of features selected by a GA to recognize depression patterns.

For NN, we chose Sigmoid as activation function to build hidden layer of size 50 and an output layer of four output neurons, representing the four depression levels (none, mild, moderate, severe). All NNs were fully connected and trained with the Adam optimizer using backpropagation with the Cross-Entropy loss function. Training a classifier on random splits of data is not adequate in human datasets, so applying leave-one-participant-out for each run then getting the average final result is a better mean for cross-validation. That is physiological data from one observer was used as the testing set, and those from the remaining participants formed the training set and repeated for all. Casper NN was used Cascade network algorithm employing Progressive RPROP to train our second neural network model which is constructed with 200 epochs, the batch size equals to 32, learning rate is 0.1 and weight decay is 0.0001. When a new neuron is added, the initial learning rates for the weights in the network are reset to values that depend on the position of the weight in the network, hence, the values of L1, L2, and L3 were set to 0.2, 0.005, and 0.001 respectively according to the paper - "A Cascade Network Algorithm Employing Progressive RPROP" (N.K. Treadgold and T.D. Gedeon, 1997). The third and fourth models were applied GA for feature selection on NN model and Casper model, respectively to compare the prediction accuracies among all models. The procedure would be taking the output bit string vector of training GA to extract the corresponding index of the features, then input these selected features to re-train the NN model and Casper NN model for depression level prediction.

E. Evaluation Measurements

To validate the effectiveness of our models, we chose precision and recall as the evaluation measurements. Precision (also called positive predictive value) is defined as the proportion of individuals that are correctly predicted with depression level (say L) actually have the depression, while recall (also known as sensitivity) is the percentage of depressed individuals that are correctly predicted with depression level L among all individuals labelled with depression

level L. Since we are predicting multiclass depression labels, we calculated the average precision, recall for all depression levels to evaluate the general prediction performance for comparisons with all models. Moreover, we also computed the overall accuracy to evaluate the overall performance, which is the number of individuals correctly predicted with their corresponding depression levels over the total number of individuals.

3 RESULTS AND DISCUSSION

I. NN Model

Since we applied leave-one-participant-out for cross-validation to build a neural network classifier with one hidden layer. At first, physiological data from one observer was used as the testing set, and those from the remaining participants formed the training set, and repeated for all 12 observers, averaging to get the final results. The overall accuracy aims to evaluate the overall performance, which is the number of individuals correctly predicted with their corresponding depression levels over the total number of individuals. We trained the model using individual signal as well as the combination of three signals, the accuracies were calculated based on the average results of 5 runs and the results are shown in Table 2. Overall, the testing accuracy is similar no matter for single signal or for all signals. One interesting finding is that the model with leave-p04-participant-out acquires the most accurate prediction across four conditions. Considering the average accuracy is locating between 25%-30%, the datasets from observers were not very good at consciously identifying the depression level of individuals in videos. For model evaluation, computing precision and recall as evaluation measures were also presented in Table 3. It is worth noting that when the NN was provided with all features derived from the physiological signals, the average precision, recall and accuracy across all levels were 5% lower than those of the model with single GSR feature classifier.

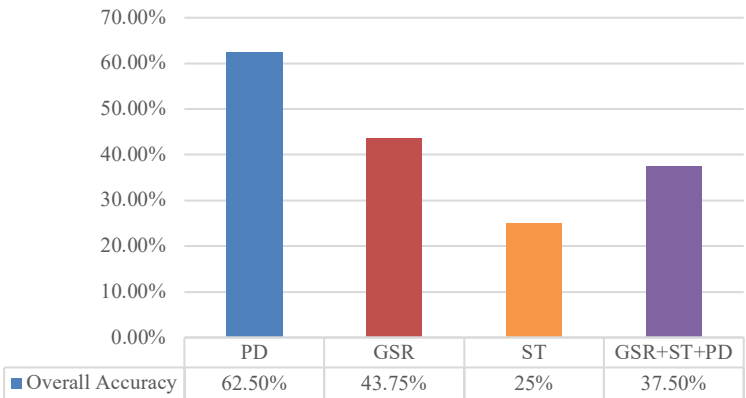
Table 2. Results of Depression Prediction from Observers' Physiological Signals from NN Model

Accuracy(%) / Signals	P02	P03	P04	P06	P07	P08	P09	P10	P11	P12	P13	P14	Average
PD	31.25	25	50	25	31.25	37.5	6.25	6.25	18.75	43.75	18.75	31.25	**25%**
GSR	18.75	37.5	43.75	25	50	37.5	18.75	18.75	25	25	37.5	25	**30%**
ST	6.25	31.25	50	25	25	31.25	37.5	37.5	31.25	18.75	31.25	31.25	**26.56%**
GSR+ST+PD	12.5	18.75	43.75	37.5	25	31.25	12.5	37.5	25	12.5	25	31.25	**24.11%**

II. Casper NN Model

After re-constructing the original NN model with Casper algorithm, in general, the testing accuracy improved a lot from Casper NN model, especially for single PD and single GSR signals, their average accuracy is 62.5% and 43.75% which is 37.5% and 13.75% more accurate than NN model, respectively. It is a slight decrease for ST signal accuracy (1.56%), and approximate 13% increase for model with all signals. It seems that Casper NN significantly solved the low testing accuracy issue with NN model; however, the precision and recall are quite low in Casper NN model (Table 3).

Fig. 4. Overall Accuracies of Depression Prediction from Single Signal and All Signals from Casper NN Model



In summary, without GA for feature selection, the specification about NN model trained with features from all signals performed slightly worse than that trained with single signal could be verified in this study. Although the predicted data is highly corresponding to the target data in Casper NN model, the number of false negatives and false positives becomes larger as well, which makes precision and recall decrease dramatically.

Table 3. Performance Measures by Depression Level for Models Defined from Single Physiological Signal

Depression Level	Signal	Precision				Recall			
		NN	Casper NN	GA+NN	GA+Casper NN	NN	Casper NN	GA+NN	GA+Casper NN
None	GSR	0.90	0.49	0.66	0.64	0.86	0.39	0.66	0.33
	ST	0.69	0.42	0.57	0.16	0.66	0.32	0.60	0.47
	PD	0.88	0.50	0.70	0.52	0.86	0.32	0.66	0.36
Mild	GSR	0.93	0.33	0.63	0.30	0.86	0.48	0.76	0.30
	ST	0.65	0.40	0.50	0.77	0.63	0.61	0.51	0.32
	PD	0.77	0.40	0.66	0.41	0.68	0.43	0.60	0.37
Moderate	GSR	0.80	0.41	0.66	0.07	0.95	0.30	0.58	0.38
	ST	0.69	0.38	0.55	0.16	0.75	0.20	0.46	0.35
	PD	0.81	0.42	0.66	0.20	0.80	0.52	0.64	0.33
Severe	GSR	0.99	0.47	0.80	0.39	0.93	0.48	0.78	0.44
	ST	0.65	0.48	0.55	0.32	0.64	0.52	0.62	0.41
	PD	0.75	0.37	0.61	0.36	0.86	0.39	0.75	0.44

Genetic Algorithm for Feature Selection

There are 85 independent features and 1 dependent feature (depr_label) existing in our dataset, and we applied genetic algorithm to select only the most descriptive features for the classification model. One of the possible subset features selected by GA was composed of 46 features and the candidate chromosomes looks like [0 1 0 0 1 1 0 1 1 0 1 0 1 1 0 1 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 1 1 0 1 0 0 0 1 1 1 0 1 1 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 1 1 0 0 1 0 1 1 1 1 1 1 0 1 0 0] (refer to Appendix 1 for corresponding extracted features) and the variation of cost function with respect to iteration as well as standard deviation of cost against iteration are plot in Figure 5 and Figure 6. We could find that the variation of cost is improved after 4000 (2/5) iterations then reached a plateau, while standard deviation of cost becomes very small after 1/5 iterations.

Fig. 5. Variation of cost against iterations

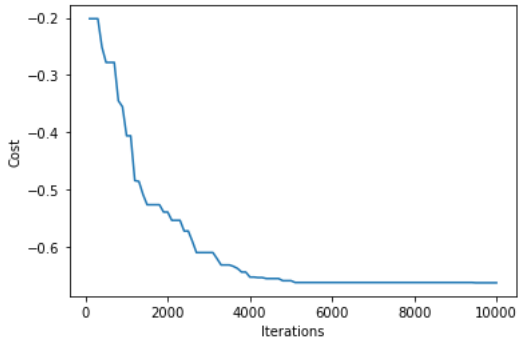
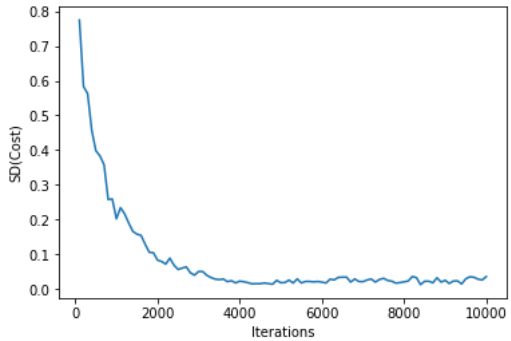


Fig. 6. Standard deviation of cost against iterations



Although we were interested in recognizing individuals' depression levels achieved by a combination of GSR, ST, and PD measurements as monitored signals. Assessment of the overall usefulness of each signal is also important as fewer sensors are required if only a single signal is needed to achieve the similar prediction accuracy that is helpful for further experimental design as well as improves the efficiency of data collection. Therefore, we applied GA for each single physiological signal to obtain the best subset of selected features, and the classifier was retrained and retest using the same validation scheme. In the end, there are 46 features were selected out of 85 depression features from GA; 21 features selected out of 39 PD features; 13 features selected out of 24 GSR features; and 11 features selected out of 39 ST features (Table 4) (refer to Appendix 2, 3, 4 for each corresponding signal with extracted features).

Table 4. Feature Reduction Rate with GA Selection

Signal	GA Features	Total Features	Reduction Rate
PD	21	39	46.15%
GSR	13	24	45.83%
ST	11	39	71.79%
GSR+ST+PD	46	85	45.88%

III. GA + NN Model

Performances of the classifications were calculated based on the average results of 5 runs for each signal and all signals. Form the comparison bar plots with NN model in Figure 7, applying GA for feature selection has slightly higher accuracy (around 3%) for single PD, single ST and all signals, while a slight decrease for single GSR signal. The more detailed result with cross-validation are also recorded in Table 5. Hence, it can be concluded that the subsets features determined by the genetic algorithm achieved similar accuracies to the models generated using all features, but with the advantage that there is an 45% to 71% reduction in the dataset size (Table 4), including all its implications, such as a faster data processing and more representative feature collections, which is very important when dealing with original experimental data since its resources are limited and expensive.

Table 5. Results of Depression Prediction from Observers' Physiological Signals from GA+NN Model

Accuracy(%) / Signals	P02	P03	P04	P06	P07	P08	P09	P10	P11	P12	P13	P14	Average
PD	31.25	18.75	56.25	31.25	25	43.75	12.5	18.75	18.75	31.25	25	31.25	28.65%
GSR	25	18.75	37.5	37.5	50	12.5	37.5	31.25	25	31.25	12.5	25	28.65%
ST	6.25	50	18.75	37.5	31.25	31.25	12.5	37.5	37.5	18.75	43.75	31.25	29.69%
GSR+ST+PD	31.25	31.25	31.25	43.75	25	12.5	6.25	25	37.5	31.25	18.75	12.5	27.60%

Fig. 7. Average Accuracies of Depression Prediction between NN Model and GA+NN Model

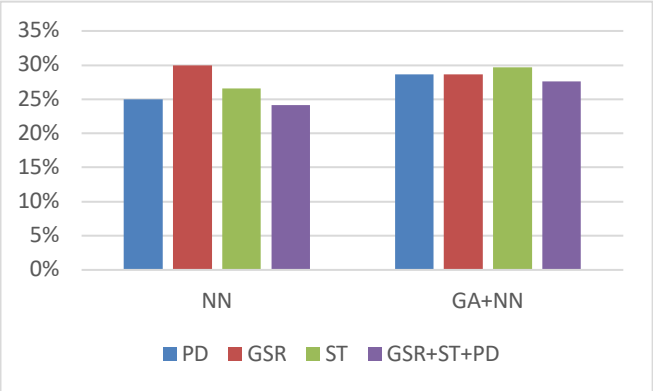
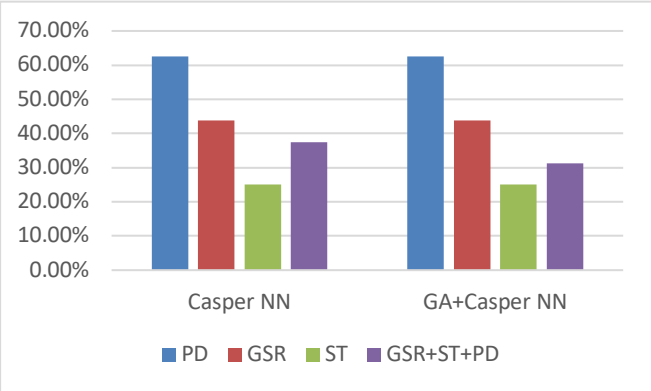


Fig. 8. Average Accuracies of Depression Prediction between Casper NN Model and GA+Casper NN Model



IV. GA + Casper NN Model

Compared the overall model prediction accuracies between with and without GA for feature selection of Casper NN models, the differences are too small to be ignored (Figure 8). So that we know GA is useful for both models to achieve the same or even higher accuracy with less input features. Nevertheless, the precision and recall values varies a lot depends on individual signal and depression level (Table 3), for example, the precision of ST signal in mild depression level increases from 0.4 to 0.77; while the precision drops from 0.4 to 0.16 for ST signal in none depression level. The same phenomenon can be seen in recall due to their trade-off property. We have tried to adjust every parameter including

number of epochs, batch size, learning rate and weight decay, the number of precision and recall do not improve accordingly. The main reason may because we only have few responses in our datasets, 176 responses in training data and 16 responses in test data. Therefore, the Casper NN model is quite complicated but built on too few data making it become inaccurate when predicting new coming test data.

In conclusion, when GA is not used for feature selection, an NN model trained with features from all signals performed slightly worse than that trained with single signal. This shows that the NN is susceptible to features that were irrelevant and redundant for depression classification. The better performance of GA+NN trained with all signals confirmed this. On the other hand, for Casper NN model, the overall accuracy has already improved compared to NN model without GA, and the optimization of GA is not that significant as NN model. Nevertheless, the GA approach demonstrates an important feature reduction, only around 30%-50% of original features reaching higher or equivalent accuracy than the models created with more features. Although genetic algorithms might be costly in computational terms since the evaluation of each individual requires the training of a model, they usually perform better than traditional feature selection techniques. In addition, genetic algorithms can manage data sets with many features, and they don't need specific knowledge about the problem under study. These advantages are proven in this study.

To further see the contribution of each signal to the prediction of each depression level, precision and recall were calculated for four models and recorded in Table 3. PD has an overall great performance on recognizing all depression levels except for “Severe” category, achieving better precision and recall on average. GSR gave the best result in identifying individuals with severe depression levels indicated by its highest performances in both measures for NN, GA+NN and GA+Casper NN models. This shows that when only one physiological signal is available, GSR is better at distinguishing the people with severe depression level, while PD is more useful in identifying healthy people and individuals with relatively mild or moderate depression levels.

4 CONCLUSION AND FUTURE WORK

The present work allows us to demonstrate that it is possible to generate depression level classification models using observers’ physiological features including GSR, ST and PD by applying feature selection through genetic algorithms on neural network model and Casper neural network model, preserving accuracies greater than 62.5% and reducing the set of data used by 45% to 72%. It is worth noting that the best accuracy in prediction achieving by single PD signal with Casper NN model, showing that observers’ PD itself can be an effective signal for predicting other individuals’ depression level while GSR and ST may provide less informative features. In addition, we also prove that GA is an important technique to obtain a smaller but more significant feature set. The classifiers with consideration of GA-based feature selection were able to eliminate the irrelevant and redundant features and meanwhile the classifiers achieved better classification performances using only fewer features.

Although the proposed physiological signals prediction approaches accomplished well detection of the observers’ depression level, there are some limitations in this study and some aspects can be worked on in the future. The proposed future work including:

I. To work with a different set of features. To obtain a larger feature set from the physiological signals like Heart Rate Variability to verify the behavior of the feature selection methods now analyzed with a different set of features and contrast the results with those obtained in this work to finally determine the subset of features that best predict depression levels.
II. To combine feature selection methods. Generate classification models using subsets of features resulting from the combination of the method proposed in the present work with the classical methods of feature selection such as forward selection and backward elimination.
III. Feature selection using GA may require a longer time. Therefore, other faster EAs can be utilized as a feature selection technique with neural network to produce a more efficient performance of the depression level detection.

Hopefully, the findings in this filed could assist current depression diagnosis with more objective measurements, which combined with the use of known effective treatments would decrease the burden for individuals and society and also have higher possibility to recover mental health in earlier stage.

5 REFERENCES

1. X. Zhu, T. Gedeon, S. Caldwell and R. Jones. (2019) Detecting emotional reactions to videos of depression, 2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES), pp. 000147-000152, doi: 10.1109/INES46365.2019.9109519.
2. Treadgold N.K., Gedeon T.D. (1997) A cascade network algorithm employing Progressive RPROP. In: Mira J., Moreno-Díaz R., Cabestany J. (eds) Biological and Artificial Computation: From Neuroscience to Technology. IWANN 1997.
3. Treadgold, N.K. & Gedeon, Tom. (1998) Exploring architecture variations in constructive cascade networks. 1. 343 - 348 vol.1. 10.1109/IJCNN.1998.682289.
4. N. K. Treadgold and T. D. Gedeon. (1999) "Exploring constructive cascade networks," in IEEE Transactions on Neural Networks, vol. 10, no. 6, pp. 1335-1350, Nov. 1999, doi: 10.1109/72.809079.
5. Khoo S., Gedeon T. (2009) Generalisation Performance vs. Architecture Variations in Constructive Cascade Networks. In: Köppen M., Kasabov N., Coghill G. (eds) Advances in Neuro-Information Processing. ICONIP 2008.
6. Hand, D., Christen, P. (2018) A note on using the F-measure for evaluating record linkage algorithms. Stat Comput 28, 539–547.
7. Chehao2628. (2020) ImprovedCasper. San Francisco (CA): GitHub; [accessed 2021 April 1]. https://github.com/chehao2628/ImprovedCasper.git
8. N. Sharma and T. Gedeon. (2013) Modeling stress recognition in typical virtual environments," in Proceedings of the 7th international conference on pervasive computing technologies for healthcare, pp. 17–24.
9. Sabah Sayed, Mohammad Nassef, Amr Badr, Ibrahim Farag. (2019) A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets, Expert Systems with Applications, Volume 121, Pages 233-243, ISSN 0957-4174
10. Garcı´a-Dominguez, A., Galva´n-Tejada, C. E., Zanella-Calzada, L. A., Gamboa-Rosales, H., Galva´n-Tejada, J. I., Celaya-Padilla, J. ´. M., Luna-Garc´ıa, H., & Magallanes-Quintanar, R. (2020, February 13) Feature Selection Using Genetic Algorithms for the Generation of a Recognition and Classification of Children Activities Model Using Environmental Sound. Hindawi. https://www.hindawi.com/journals/misy/2020/8617430/#materials-and-methods
11. Tong, D.L., Mintram, R. (2010) Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int. J. Mach. Learn. & Cyber. 1, 75–87. https://doi-org.virtual.anu.edu.au/10.1007/s13042-010-0004-x
12. Prakhar Gurawa. (2021) Feature-Selection-Using-Genetic-Algorithm. San Francisco (CA): GitHub; [accessed 2021 May 1]. https://github.com/prakhargurawa/Feature-Selection-Using-Genetic-Algorithm.git
13. Waleed Ali, Adel A. Ahmed. (2019) Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. The Institution of Engineering and Technology.https://doi-org.virtual.anu.edu.au/10.1049/iet-ifs.2019.0006

6 APPENDIX

Appendix 1: GA Feature Selection for All Depression Features

Depression Features	Bit Value
max_normalised_skintemp, mean_normalised_skintemp, std_normalised_skintemp, var_normalised_skintemp, rms_normalised_skintemp, var_filtered_skintemp, first_diff_filtered_skintemp_abs_mean, second_diff_filtered_skintemp_abs_mean, num_normalised_skintemp_peaks, vlp_skintemp_peak_occurrences, lp_skintemp_peak_occurrences, mean_vlp_skintemp_peak_amplitudes, min_normalised_pupil_left, max_normalised_pupil_left, mean_normalised_pupil_left, std_normalised_pupil_left, rms_normalised_pupil_left, vlp_pd_left_peak_occurrences, lp_pd_left_peak_occurrences, mean_vlp_pd_left_peak_amplitudes, mean_lp_pd_left_peak_amplitudes, ratio_peak_occurrence_vlp_lp_left, min_normalised_pupil_right, max_normalised_pupil_right, mean_normalised_pupil_right, std_normalised_pupil_right, first_diff_normalised_pupil_right_abs_mean, second_diff_normalised_pupil_right_abs_mean, mean_lp_pd_right_peak_amplitudes, ratio_peak_occurrence_vlp_lp_right, max_normalised_pupil_avg, mean_normalised_pupil_avg, var_normalised_pupil_avg, rms_normalised_pupil_avg, lp_pd_avg_peak_occurrences, mean_vlp_pd_avg_peak_amplitudes, rms_normalised_gsr, max_filtered_gsr, mean_filtered_gsr, var_filtered_gsr, rms_filtered_gsr, first_diff_filtered_gsr_abs_mean, second_diff_filtered_gsr_abs_mean, num_filtered_gsr_peaks, vlp_sc_scr_occurrences, mean_vlp_sc_scr_amplitudes	1
min_normalised_skintemp, min_filtered_skintemp, max_filtered_skintemp, mean_filtered_skintemp, std_filtered_skintemp, rms_filtered_skintemp, first_diff_normalised_skintemp_abs_mean, second_diff_normalised_skintemp_abs_mean, num_filtered_skintemp_peaks, mean_lp_skintemp_peak_amplitudes, ratio_skintemp_peak_occurrence_vlp_lp, var_normalised_pupil_left, first_diff_normalised_pupil_left_abs_mean, second_diff_normalised_pupil_left_abs_mean, var_normalised_pupil_right, rms_normalised_pupil_right,	0

vlp_pd_right_peak_occurrences, lp_pd_right_peak_occurrences, mean_vlp_pd_right_peak_amplitudes, min_normalised_pupil_avg, std_normalised_pupil_avg, first_diff_normalised_pupil_avg_abs_mean, second_diff_normalised_pupil_avg_abs_mean, vlp_pd_avg_peak_occurrences, mean_lp_pd_avg_peak_amplitudes, ratio_peak_occurrence_vlp_lp_avg, min_normalised_gsr, max_normalised_gsr, mean_normalised_gsr, std_normalised_gsr, var_normalised_gsr, min_filtered_gsr, std_filtered_gsr, first_diff_normalised_gsr_abs_mean, second_diff_normalised_gsr_abs_mean, num_normalised_gsr_peaks, lp_sc_scr_occurrences, mean_lp_sc_scr_amplitudes, ratio_scr_occurrence_vlp_lp	

Appendix 2: GA Feature Selection for PD Features

PD Features	Bit Value
min_normalised_pupil_left, mean_normalised_pupil_left, std_normalised_pupil_left, rms_normalised_pupil_left, vlp_pd_left_peak_occurrences, lp_pd_left_peak_occurrences, mean_vlp_pd_left_peak_amplitudes, mean_lp_pd_left_peak_amplitudes, ratio_peak_occurrence_vlp_lp_left, min_normalised_pupil_right, var_normalised_pupil_right, first_diff_normalised_pupil_right_abs_mean, second_diff_normalised_pupil_right_abs_mean, ratio_peak_occurrence_vlp_lp_right, max_normalised_pupil_avg, mean_normalised_pupil_avg, var_normalised_pupil_avg, rms_normalised_pupil_avg, vlp_pd_avg_peak_occurrences, lp_pd_avg_peak_occurrences, ratio_peak_occurrence_vlp_lp_avg	1

Appendix 3: GA Feature Selection for GSR Features

GSR Features	Bit Value
mean_normalised_gsr, mean_filtered_gsr, std_filtered_gsr, var_filtered_gsr, rms_filtered_gsr, first_diff_normalised_gsr_abs_mean, first_diff_filtered_gsr_abs_mean, second_diff_filtered_gsr_abs_mean, num_normalised_gsr_peaks, num_filtered_gsr_peaks, vlp_sc_scr_occurrences, mean_vlp_sc_scr_amplitudes, mean_lp_sc_scr_amplitudes	1

Appendix 4: GA Feature Selection for ST Features

ST Features	Bit Value
min_normalised_skintemp, max_normalised_skintemp, mean_normalised_skintemp, var_normalised_skintemp, rms_normalised_skintemp, max_filtered_skintemp, rms_filtered_skintemp, second_diff_normalised_skintemp_abs_mean, second_diff_filtered_skintemp_abs_mean, num_normalised_skintemp_peaks, ratio_skintemp_peak_occurrence_vlp_lp	1