Automatic Depression Detection by Multi-model Ensemble

Koh Hui Wen 1003593 | Loh De Rong 1003557 | Ma Teck Leck 1003803 | Glenda Wee Qihui 1003903

December 2020

1 Abstract

Depression is a common mood disorder that affects more than 264 million people worldwide [1]. With depression manifesting itself in many ways with different people, a diagnosis for depression is subject to interpretation by the clinician. In supporting the move toward more objective detection methods, our team proposes a multi-model ensemble approach for Automatic Depression Detection (ADD) in patients. In this process, features from audio, text and gaze datasets were fed into binary classifiers such as logistic regression and random forest. Using repeated K-cross validation, models with F1-score of greater than 0.60 were shortlisted for ensemble learning. The best combination of models with the highest validation F1-score was then selected as our final ensemble. It achieved a F1-score of 0.70 on the validation set and 0.50 on the test set. The source code can be found on GitHub.

2 Introduction

2.1 Problem Statement

The National Institute of Mental Health has stated that major depression is one of the most common mental health disorders in the United States [2]. If left untreated, this disorder can greatly affect one's physical health and well-being, which then brings about negative impacts to the people around them as well. Approximately 60% of suicide cases come from people with mood disorders like depression [3]. This alarming number creates a need for us to be able to detect and treat mental disorders like depression as soon as possible in order to give aid to those who are in need.

However, current traditional clinical methods of diagnosing depression depends on the clinician, with the accuracy of depression detection in different countries varying greatly, even in developed nations [4]. The most accurate clinicians were from the Netherlands at 83.5% accuracy, while the UK had the lowest with 45.6% accuracy. This large range of accuracies proves that there is a need for an objective depression detection strategy that is not subject to the interpretations of different clinicians.

Therefore, our project aims to explore Automatic Depression Detection (ADD), which could be a possible option to manage this mental disorder, where computer algorithms are used to aid clinicians in the detection of depression in a patient. Classification models were trained on the English-speakers database Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ). We propose an ensemble system that encompasses several models that were trained using different data mediums including audio, transcript files.

2.2 About the Dataset

To measure the severity of depression, DAIC was constructed [5]. It contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety,

depression, and post-traumatic stress disorder. The interviews were conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room, which contributes to the larger effort of creating a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness [6]. The data collected has been transcribed and annotated for a variety of verbal and non-verbal features.

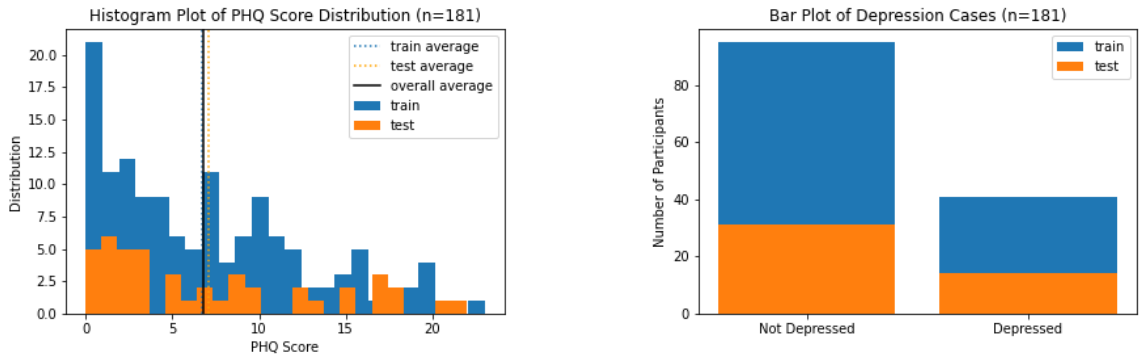
The provided Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database, which is part of the larger corpus DAIC, includes transcripts, audio recordings, audio and non-verbal features. It also contains depression scores of each participant based on the eight-item Patient Health Questionnaire depression scale (PHQ-8), which is established as a valid diagnostic and severity measure for depressive disorders in large clinical studies [7]. The total score ranges from 0 to 24 points. Following the guidance of the Depression Classification Sub-Challenge (DCC), participants with a PHQ-8 score of 10 points and above will be labeled as depressed.

For the purpose of our investigation, we used the verbal cues such as the interview audio file and the transcript of the interview. Audio files include the entire dialog between Ellie and each participant. In addition to recording the dialogue between Ellie and the participant, the transcripts also include the actions of the participants during the interview, such as sighing and laughing. For non verbal cues, gaze was also used where the participant's gaze direction was recorded throughout the entire interview.

3 Approach

3.1 Preliminary Analysis of DAIC-WOZ Dataset

Data cleaning was first performed to remove 8 participants from the dataset because their sessions either had been excluded, interrupted or contained missing transcripts. The resulting dataset consists of a total of 181 participants, of which 45 belong to the given test set.



(a) By PHQ-8 scores

(b) By binary class labels

Figure 1: Distribution of DAIC-WOZ dataset

The DAIC-WOZ dataset clearly poses two challenges to machine learning tasks. First, the number of available samples is small, so the chances of overfitting are considerably higher. Second, the dataset is highly imbalanced as there are more than twice the number of non-depressed individuals than depressed individuals, which will likely lead to inaccurate predictions that are biased toward the majority. The methods on how we mitigate these two problems are further elaborated in Section 3.2.

Supervised learning is naturally the popular approach since the labels for this dataset have been provided. For this depression task, many researchers have reported promising results using regression and classification models [8, 9].

However, our team has decided against using regression models due to reservations about the uneven data distribution in the train and test set (Figure 1a). Coupled with the fact that there are very minimal depressed features to learn from due to the small dataset, it is very easy for the regression model to overfit to the majority class in order to reduce training error. Additionally, the given test set has a similar uneven distribution as the train set. This means that any overfitted model will be able to produce small mean absolute test error as long as it is close to the average PHQ-8 scores of the train set. The consequence is that most of the depressed individuals will be incorrectly predicted with low scores, when in fact they have very high scores that require immediate attention. In this regard, we will be using binary classification models to take advantage of the larger aggregated number present in each class (Figure 1b).

3.2 Evaluation Method

3.2.1 Evaluation Metric

For a depression detection task, the true positive cases require more attention and it is more important to correctly identify them for treatment. In other words, the goal is for the model to have high true positive rates and low false negative rates (i.e. high recall score). However, focusing only on the recall score is not practical in a small dataset as it likely leads to an overfitted model that predicts every individual as depressed, resulting in increased false positive rates.

Hence during model evaluation, we used binary F1-scores as the key metric as it still places the emphasis on correctly identifying true positive cases while balancing both the precision and recall scores. The model with the highest binary F1-score, which is obtained using the k-cross validation technique, will be used for the final prediction on the test set.

3.2.2 K-cross Validation

In order to estimate how well our model performed in general on unseen data, we used k-cross validation and recorded the mean F1-scores across all iterations. K-cross validation randomly splits the training data into k folds, and at each of the k iterations, k-1 folds are used for training while one fold is used for validation. This ensures our model is not biased toward any particular validation set, hence improving its generalization.

Initially, we chose the value of k to be 10, but we realized that due to our small dataset, setting a large value for k would result in a very small validation set at each iteration of the k-cross validation. It is very likely that the validation set was highly imbalanced, thus resulting in higher performance if the model were to predict all positives or all negatives. This is extremely flawed as it misled us to believe that the model was performing well when it was not. Thus, we decreased the value of k to 4, but also repeated the process of k-cross validation 3 times for each classifier. This repeated k-cross validation allows us to reduce the error in estimating the mean model performance and improve generalization.

3.2.3 Undersampling

As the dataset is imbalanced, classifiers may be biased toward predicting the negative case. Thus, we used undersampling to make the dataset balanced, with 50% of the samples being depressed and 50% of the samples being non-depressed. With reference to the existing training and test set that is given by the DAIC-WOZ database, our team also followed the same grouping for our own training and test sets, ensuring consistency across all datasets. Shuffling with the same random seed on the training data points was also performed to minimise bias from points before them. After resampling, the sample size of the train set was reduced to 82 (41 negative cases, 41 positive cases).

Comparing the mean validation scores for the original dataset and the undersampled dataset, we see that the undersampled dataset consistently performed better (Appendix A). Hence, we would be referring to the undersampled dataset for the rest of this paper unless otherwise mentioned.

3.3 Proposed Workflow

Each of the classifier's hyperparameters was first fine-tuned by taking the best F1-score from repeated k-cross validation. Following which, only those classifiers with validation F1-scores greater than 0.60 were shortlisted to be used in our ensemble (Figure 2). The ensemble takes in the probability output of each classifier and averages these probabilities to produce the final probability of the sample having depression (Figure 3). If the final probability is more than 0.5, the sample is predicted as depressed. Otherwise, it is predicted as non-depressed.

The intention of using an ensemble is to improve the robustness of our model, as it reduces the bias that some classifiers have toward certain features, improving the performance of the output. This is because an ensemble reduces the spread or dispersion of the predictions and model performance. As the performance of the individual classifiers also affect the performance of the ensemble, the input models must also achieve a certain level of performance. We chose to only consider input classifiers that produce a mean validation F1-score of at least 0.60 during k-cross validation.

We also made further experiments, testing out different combinations of input classifiers. Repeated k-cross validation was used on our ensemble, and the ensemble with the highest mean validation F1-score was chosen to be our final model.

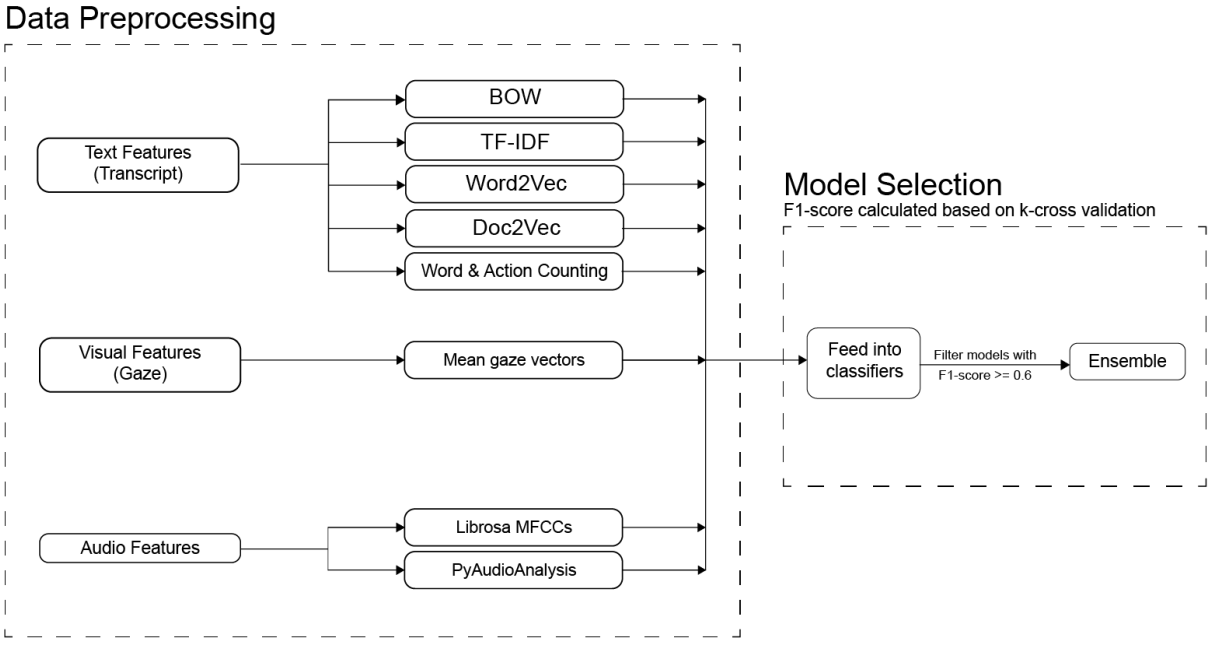


Figure 2: Overall Project Workflow

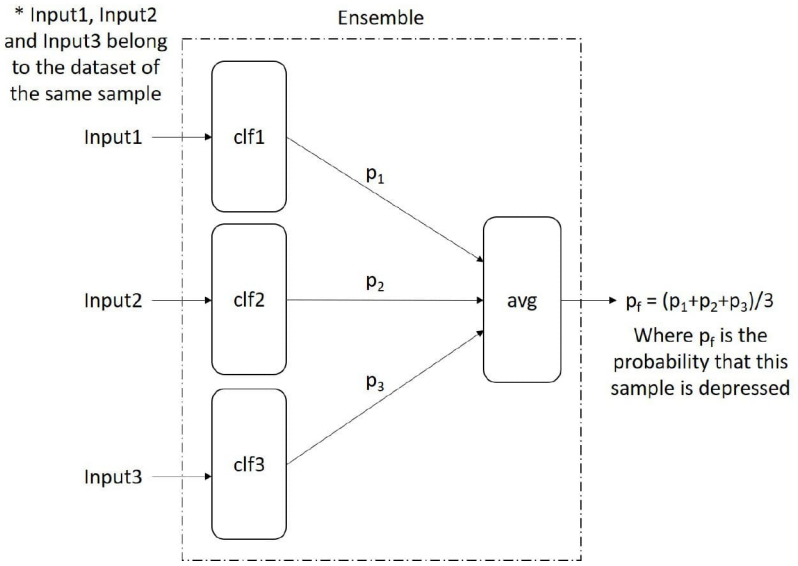


Figure 3: Ensemble of 3 Classifiers

4 Data Pre-processing

4.1 Text Features

Clear and consistent differences in the content and style of language have been noted between those with and without symptoms of depression [10]. For example, depressed individuals may use an excessive amount of words conveying negative emotions such as “sad” and “lonely”, or use more absolutist words such as “nothing” and completely”. Hence, to tackle the depression detection task, extracting appropriate text features from the transcript becomes important.

Each transcript file contains lines, marked by timestamps, that are spoken by the participant and by the interviewer Ellie. In the first step of data cleaning, each participant’s words were first concatenated into a single text document, followed by a compilation of all the documents that matched their corresponding class labels into a single dataframe.

Text wrangling was then performed. First, a blacklist containing common english stop words was initialized, in which personal pronouns were intentionally excluded. As there have been several articles that show that depressed persons tend to use first-person pronouns more (e.g. ”I”, ”me”, ”myself”), we speculated the comparison of the usage in personal pronouns could produce several important features [11]. To this end, we separately extracted 13 features from the raw transcripts which are thought to be linked to depression. They represent the frequency of the participants’ actions as well as the pronouns and absolutist words. Second, the sentences in each document were then processed through text normalization using contractions library, tokenization using TweetTokenizer and the removal of HTML tags, punctuations, blacklisted words. The final output was a clean compiled transcript, which allowed for further separate processing to obtain higher order features (Table 1).

Table 1: Summary of pre-processing steps for text features

Features	**Remarks on additional pre-processing steps**
Bag of Words (BOW)	CountVectorizer was used. Words that appeared less than 3 times were not included in the occurrence matrix.
Term Frequency-Inverse Document Frequency (TF-IDF)	TfidfVectorizer was used. As the matrix was very sparse, the models were trained on the dense matrix representation of the vectorized transcripts.
Document embeddings (word level) - Word2Vec	Each document is represented by mean pooling. The vectors obtained from the Word2Vec weights trained on Google News corpus. Dimension = 300.
Document embeddings (document level) - Doc2Vec	Dimension = 300. It extends the idea of Word2Vec and provides the relationships between documents and not just words.
Word and Action Counting	For each document, the frequencies of these 13 features were counted. Actions: laughs, sighs, throat clears, yawns, sniffs, coughs, sharp inhales, sneezes, deep breaths. Significant vocabularies: singular first-person pronouns, plural first-person pronouns, third-person pronouns, absolutist words.

4.2 Visual Features

The dataset provided includes several visual features, such as the 3D points on the face, gaze direction, and pose. Due to limited time, we chose to focus on the gaze direction, as research shows that the gaze may be correlated to mental illnesses [12].

Each gaze file contains 4 vectors, accompanied by the success and confidence level of detection, at each periodic interval throughout the interview. The first two vectors, $f0 = (x_0, y_0, z_0)$ and $f1 = (x_1, y_1, z_1)$, are in world coordinate space and describes the gaze direction of both eyes, while the second two vectors, $fh0 = (x_h0, y_h0, z_h0)$ and $fh1 = (x_h1, y_h1, z_h1)$, describe the gaze in head coordinate space [5].

Since not all gaze features were detected with success, we filtered out the frames with success being 0 and confidence level below a threshold T. With larger values of T, the gaze features would be more accurate, reducing the number of outliers or errors. However, some participants have lower confidence levels in gaze detection, and the largest possible threshold T we could use without removing too much data is 0.90.

After the filtering process, we took the mean of each vector across the timestamps for each participant to represent the participants' overall gaze direction. The resulting gaze features for each participant consisted of 12 values, or 4 vectors. The gaze features for each participant is then appended into a 2D array, where one row is one sample and one column is one of the 12 values. During training of classifiers, we also further broke down these 12 values into the 4 vectors by slicing the array by columns.

4.3 Audio Features

Noticeable differences in speech production between depressed and non-depressed patients have been suggested as a potential biomarker [13]. A person's speech can measure the severity of depression. For example, the speech of depressed patients changes when they respond to certain questions, becoming faster and with shorter pauses. Therefore, for our depression detection task, we decided to extract useful audio features from the DAIC-WOZ dataset as well.

The audio recordings had to be pre-processed and cleaned because they contained periods where both the participant and Ellie remained silent, and some have heavy static noises. First, we intended to remove silent segments from the audio files by segmenting each audio file into 1-second audio files and then removing segments whose computed energy falls below a threshold of 50% of the median energy value. This procedure removed segments with relatively lower volume for each audio recording. The silence removal procedure also managed to remove Ellie's voice for most of the recordings as the participant's voice was relatively louder due to the usage of a head mounted microphone during the interview.

However, this process did not remove heavy static noises. We attempted to use a noise reduction library [14], but this process instead produced audio files that had even more silent periods than the unprocessed audio files. In the interest of time, we removed the recordings that consist of heavy static noises to ensure that the training data is clean.

With the clean audio data, we used Librosa to extract Mel Frequency Cepstral Coefficients (MFCCs) which are state-of-the-art audio features to learn about an audio sample [15]. We also used pyAudioAnalysis to extract 68 short-term audio features for each segment [16], and then performed mean pooling for all the segments that eventually represented each audio sample.

5 Results and Discussion

5.1 Mean Validation F1-scores of Individual Models

For each feature dataset, we ran a grid search to find the best parameters for 4 different types of models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). The mean k-cross validation F1-score for each tuned model was recorded.

Table 2: Mean validation F1-scores of all classifiers

Features	**Classifiers**			
	LR	**DT**	**RF**	**SVM**
BOW	**0.6440**	**0.6473**	**0.6121**	0.5547
TF-IDF	0.2579	**0.6121**	0.5502	**0.6976**
Word2Vec	**0.6225**	**0.6538**	0.5773	**0.6041**
Doc2Vec	0.5147	0.5478	0.5993	0.5830
Word and Action Counting	0.3929	**0.6151**	0.5614	0.5810
Gaze (all vectors)	0.3702	0.5398	0.5169	0.5379
Gaze (f0)	0.4165	0.5506	0.4621	0.4832
Gaze (f1)	0.3935	**0.6003**	0.5832	0.5537
Gaze (fh0)	0.3365	0.5219	0.4658	0.4181
Gaze (fh1)	0.3844	**0.6482**	0.5233	0.4746
Librosa (MFCCs)	0.2300	0.4411	0.3202	0.2305
PyAudioAnalysis	0.2847	0.3322	0.2314	0.2033

5.2 Multi-model Ensemble

The input classifiers to our ensemble learning model were chosen from the models in Table 2, and we only considered the models that produce mean validation F1-score of at least 0.60 (bolded). We then performed repeated k-cross validation on each ensemble, recording the mean F1-score (Appendix B). The best performing model is shown in Table 3.

Table 3: Best performing ensemble based on validation F1-score

Features	**Classifiers**				**F1-score**
	LR	**DT**	**RF**	**SVM**	
BOW	✓	✓	✓		0.70
Word2Vec	✓	✓		✓	
TF-IDF				✓	
Word and Action Counting		✓			
Gaze (f1)		✓			
Gaze (fh1)		✓			

5.3 Evaluation Against Other Models

With our best ensemble which has the highest mean validation score, we evaluated results on the DAIC-WOZ test set, and compared them with the baseline SVM-based classifier provided by the AVEC-2016 challenge and the models used by Vázquez-Romero and Gallardo-Antolín [17]. These metrics are shown for both the depressed and non-depressed classes (in parentheses). However, we have to keep in mind that the comparison below only provides an estimate as our test set has 2 less participants than the one provided by DAIC-WOZ due to missing data as mentioned in Section 3.1.

Table 4: Performance on test set compared with other research

Method	**F1-score**	**Precision**	**Recall**
Baseline (SVM)	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)
Ensemble 50 1d-CNN	0.65 (0.80)	0.55 (0.89)	0.79 (0.73)
Our Multi-model Ensemble	0.50 (0.60)	0.38 (0.79)	0.71 (0.48)

The SVM-based classifier uses hand-crafted speech features extracted with COVAREP (v1.3.2), an open-source MATLAB and Octave toolbox [18] while our final ensemble uses a combination of text, visual and handcrafted features. The presented Ensemble 50 1d-Convolutional Neural Network (CNN) model, on the other hand, adopted a refined 1d-CNN architecture based on a CNN + Long Short-Term Memory (LSTM)-based system called DepAudioNet [19]. Even though the authors only explored audio features using speech log-spectrogram, their proposed model was able to capture frequency correlations at short-term level through the use of 1d convolutional layer and temporal dynamics through the use of 1d pooling layer.

Referring to Table 4, our ensemble performs slightly better than the baseline model in terms of F1-score, but worse than the model used in the compared research. The reason why our ensemble and the baseline model yielded similarly poor results is likely because of our naive model implementation. Despite taking more features into account for our ensemble, we did not consider the semantic relationships between features.

The slight increase in performance of our model as compared to the baseline model could be because we used repeated k-cross validation to help us fine-tune the hyperparameters during model training as well as select the best models for the final ensemble. This similar process of k-cross validation was also reflected in the compared study when choosing the most optimal number of filters and max-pooling window size.

Comparing our model's performance on the test set with the mean validation F1-score, the ensemble performed much worse on the test set. We had too little data to learn about all the general features, thus the model became overfitted. We suspect that the difference could also be due to the train and test sets not belonging to the same distribution, but we have too little data to confirm this.

6 User Interface

Figure 4 shows the user interface for our work. The input text-box can take in a test participant's ID which ranges from 0-44. Clicking on the respective buttons would yield the transcript, true label and the predicted label for the selected participant.

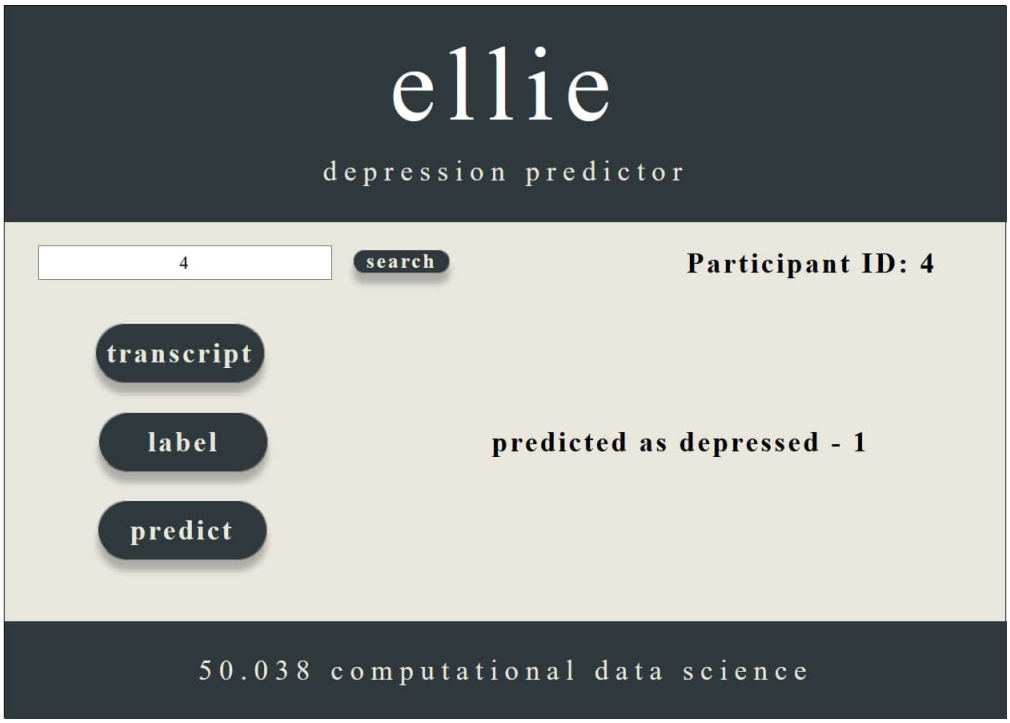


Figure 4: User Interface

7 Conclusion

As seen in Table 4, our ensemble may perform poorly compared to other existing models, thus there is definitely room for improvement. We will discuss the experiments that have failed and possible improvements that could have been made, but were not implemented due to limited time.

7.1 Failed Experiments

For each of the feature datasets, we attempted to use neural networks. Firstly, we used fully connected layers with varying numbers of units (ranging from 5 to 128). We also attempted to use LSTM with 3D sequence inputs (still containing temporal features), as well as convolutional layers. However, regardless of the neural network structure and hyperparameters, the model had a tendency to predict all samples as negative, or all samples as positive. Even at lower learning rates, the model converged after a few epochs, and the F1-score was often close to zero. Perhaps the poor performance of the neural networks stems from our very limited dataset [20] and the neural networks over-fitted very easily. However, we only used neural networks as a classifier and did not try to extract higher dimensional features to input into another model instead, which could be an area of further research.

The audio classification models did not perform well compared to the other feature datasets. Even after removing silent audio segments, the noise throughout the audio was not removed. In order to obtain a clean train set, we had to manually remove those with static noises. However,

this greatly reduced our training samples, which likely resulted in poor validation F1-score that was much worse than the cut-off score of 0.60 for the ensemble.

7.2 Future Improvements

For all the features that were fed into the classifiers of our final ensemble, the temporal aspect was not taken into account. The surrounding context was also not being considered. In future iterations, we could extract higher order temporal features through the use of contextual word embeddings such as Embeddings from Language Models (ELMo) followed by LSTM and attention models which would provide greater semantic meanings and likely better predictions.

Another approach is to use a neural conversational model, such as the Seq2Seq model, that implicitly learns the semantic and syntactic relations between pairs and captures the contextual dependencies that are not possible with conventional machine learning approaches [21]. In this manner, more meaningful features can be extracted from the exchanges between Ellie and the participants.

We could also build models for the other visual features such as the pose or 3D points on the face. Analysis of head poses and movement were shown to be effective cues when it comes to detecting depression in other studies [22], as it was seen that behavioural elements are more pronounced in the head and hand regions for depressed patients. Therefore, the exploration of other non-verbal cues could also help further improve the performance of our model.

Finally, to address the problem of a small dataset, oversampling can be considered. A more careful design of the k-cross validation can be created such that the train set can be oversampled, using the Synthetic Minority Oversampling Technique (SMOTE) for the gaze and audio features or paraphrasing techniques for the transcript features, in a way that does not bleed into the validation set for every generated fold. This allows us to expand the usage of our small dataset.

7.3 Future Implications

In conclusion, detecting depression is not an easy task. We need to dig deeper into the meaning and relationship of the extracted features, rather than simply using the raw features. In addition, the small and imbalanced dataset clearly poses a huge challenge. Although we partially mitigated this problem through the use of undersampling and k-cross validation, there is definitely still room for improvement.

References

[1] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] N. I. of Mental Health, "Major depression," accessed: 2020-12-01. [Online]. Available: https://www.nimh.nih.gov/health/statistics/major-depression.shtml

[3] U. S. D. of Health & Human Services, "Does depression increase the risk for suicide?" accessed: 2020-12-01. [Online]. Available: https://www.hhs.gov/answers/mental-health-and-substance-abuse/does-depression-increase-risk-of-suicide/index.html

[4] A. J. Mitchell, S. Rao, and A. Vaze, "International comparison of clinicians' ability to identify depression in primary care: meta-analysis and meta-regression of predictors," *British Journal of General Practice*, vol. 61, no. 583, pp. e72–e80, 2011.

[5] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[6] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[7] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[8] B. Li, J. Zhu, and C. Wang, "Depression severity prediction by multi-model fusion," in *HEALTHINFO 2018 : The Third International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*, 2018, pp. 19–24.

[9] H. Dinkel, M. Wu, and K. Yu, "Text-based depression detection: What triggers an alert," *arXiv preprint arXiv:1904.05154*, 2019.

[10] M. Al-Mosaiwi, "People with depression use language differently–here's how to spot it," *Retrieved form http://theconversation. com/people-with-depression-use-language-differently-heres-how-to-spot-it-90877*, 2018.

[11] T. Brockmeyer, J. Zimmermann, D. Kulessa, M. Hautzinger, H. Bents, H.-C. Friederich, W. Herzog, and M. Backenstrass, "Me, myself, and i: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety," *Frontiers in psychology*, vol. 6, p. 1564, 2015.

[12] R. J. Larsen and T. K. Shackelford, "Gaze avoidance: Personality and social judgments of people who avoid direct face-to-face contact," *Personality and individual differences*, vol. 21, no. 6, pp. 907–917, 1996.

[13] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[14] T. Sainberg, "noisereduce," https://github.com/timsainb/noisereduce, 2020.

[15] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[16] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[17] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, 2020.

[18] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[19] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

[20] R. Bhatia, "When not to use neural networks," 2018, accessed: 2020-12-01. [Online]. Available: https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

[21] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[22] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 283–288.

Appendices

Appendix A Undersampling vs No sampling

Table 5: Comparison between undersampling and no sampling (in parentheses)

Features	**Best model**	**Validation F1-score**
BOW	DT (LR)	0.6473 (0.4600)
Word2Vec	DT (LR)	0.6538 (0.4543)
Doc2Vec	SVM (LR)	0.5830 (0.4022)
TF-IDF	SVM (DT)	0.6976 (0.4836)
Word and Action Counting	DT (DT)	0.6687 (0.4791)
Gaze (all vectors)	DT (DT)	0.5398 (0.2861)
Gaze (f0)	DT (DT)	0.5506 (0.2731)
Gaze (f1)	DT (DT)	0.6003 (0.3226)
Gaze (fh0)	DT (DT)	0.5219 (0.3529)
Gaze (fh1)	DT (DT)	0.6482 (0.2205)

Appendix B Mean validation F1-scores of ensembles

Table 6: K-cross validation results of different ensembles

Ensemble no.	Features	Classifiers				Validation F1-score
		LR	DT	RF	SVM	
1	BOW	✓	✓	✓		0.6988
	Word2Vec	✓	✓		✓	
	TF-IDF				✓	
	Word and Action Counting					
	Gaze (f1)		✓			
	Gaze (fh1)		✓			
2	BOW	✓	✓	✓		0.6794
	Word2Vec	✓	✓		✓	
	TF-IDF				✓	
	Gaze (f1)		✓			
	Gaze (fh1)		✓			
3	BOW	✓	✓	✓		0.6601
	Word2Vec	✓	✓		✓	
	TF-IDF				✓	
	Word and Action Counting		✓			
4	BOW	✓	✓	✓		0.6486
	Word2Vec	✓	✓		✓	
	TF-IDF				✓	
5	BOW	✓	✓	✓		0.6685
	Word2Vec	✓	✓		✓	
	Gaze (f1)		✓			
	Gaze (fh1)		✓			
6	Word2Vec	✓	✓		✓	0.6496
	TF-IDF				✓	
	Word and Action Counting		✓			
	Gaze (f1)		✓			
	Gaze (fh1)		✓			
7	BOW	✓	✓	✓		0.6590
	TF-IDF				✓	
	Word and Action Counting		✓			
	Gaze (f1)		✓			
	Gaze (fh1)		✓			
8	TF-IDF				✓	0.5871
	Word and Action Counting		✓			
	Gaze (f1)		✓			
	Gaze (fh1)		✓			