National College of Ireland

Detecting Depression from Speech with Residual Learning

MSc Research Project
Data Analytics

Donovan Michael Jeremiah
Student ID: x18181562

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Donovan Michael Jeremiah
Student ID:	x18181562
Programme:	Data Analytics
Year:	2019-2020
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	17/08/2020
Project Title:	Detecting Depression from Speech with Residual Learning
Word Count:	6319
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting Depression from Speech with Residual Learning

Donovan Michael Jeremiah
x18181562

Abstract

According to the World Health Organization (WHO)[1], more than 25% of the European Union's (EU) population suffer from various levels of depression and anxiety, which if left untreated, could lead to serious health disorders such as Major Depressive Disorder (MDD) or otherwise called clinical depression. Health conditions like depression and anxiety cost the EU over €170 billion every year. This study investigates the effectiveness of residual networks in depression detection. It proposes the use of ResNet-18 to predict if an individual is depressed or not, and compares its performance to a Base-CNN and AlexNet. The models are trained on the log-scaled spectrograms of the participant audio recordings from the DAIC-WOZ dataset. Preprocessing steps such as random undersampling and k-fold cross-validation contribute significantly to the performance of the models. The ResNet-18 model provides a substantially high F1-score of 0.83 which is 7.2% higher than the next best state-of-the-art model. This research demonstrates the effectiveness of residual networks in depression detection and, hence, advocates its viable use in listening and depression helpline services. One of the limitations of the model is that it shows signs of overfitting. Future work could potentially investigate the use of General Adversarial Networks (GAN) for data augmentation techniques.

1 Introduction

According to a recent study[2] by Eurofound in 2019, more than 13% of young people (aged 18-24) are at risk of depression in Ireland (see. Figure 1). Over 12% of its young people (aged 15-24) suffer from chronic depression. While the numbers are still high, the graph shows that Ireland has done relatively better than other EU countries when it comes to diagnosing chronic depression among young people. This could be attributed to the reach of dedicated medical professionals, psychiatrists, and councillors. A recent COVID-19 mental health survey[3] by Maynooth University (MU) and Trinity College of Dublin (TCD) shows that 23% of adults in Ireland reported suffering from depression and 20% from anxiety. Hence, it is evident that the ability to effectively detect and treat

[1] https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2012/10/depression-in-europe/depression-in-europe-facts-and-figures

[2] https://www.eurofound.europa.eu/publications/report/2019/inequalities-in-the-access-of-young-people-to-information-and-support-services

[3] https://www.maynoothuniversity.ie/news-events/covid-19-mental-health-survey-maynooth-university-and-trinity-college-finds-high-rates-anxiety

individuals suffering from clinical depression is paramount to the well-being of a country's society.

While traditional methods of treatment such as Cognitive Behavioural Therapy (CBT), prescribed medication, and social interventions have been used to treat MDD, they first require the concerned individual to be identified. This is done by putting them through certain invasive diagnostic procedures that are uncomfortable for those in need of help. This is where listening services (e.g. NiteLine[4]) and other depression helplines play an important role with their non-invasive methods of identifying individuals with signs of depression.

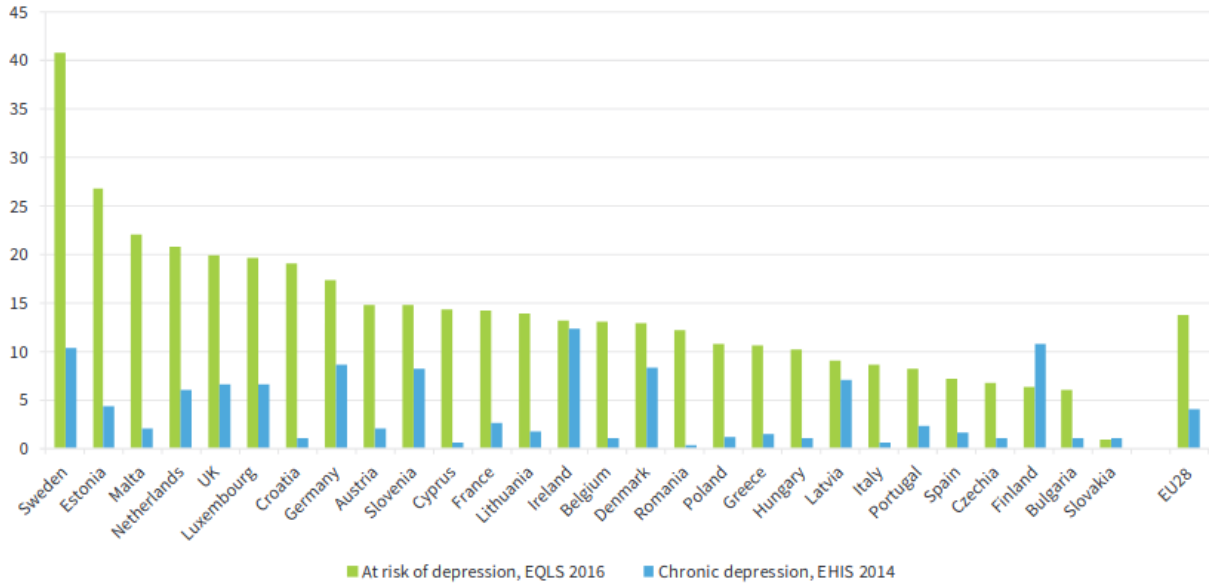


Figure 1: Risk of depression (18-24 years, 2016) and reported chronic depression (15-24 years, 2014), by country (%)

The current studies in depression detection use the Wizard-of-Oz interviews from the Distress Analysis Interview Corpus (DAIC-WOZ) to study depression among adults. Studies by Haque et al. (2018), Srimadhur and Lalitha (2020), and Yalamanchili et al. (2020), have shown the immense power of neural networks being applied to the field of depression detection. However, with the recent introduction of ResNets (He et al.; 2016a) in the computer vision domain, there has surprisingly been less research investigating the use of residual networks in depression detection. ResNets are a type of Deep Convolutional Neural Network (DCNN) that achieved incredible success in computer vision's prestigious competition ILSVRC 2015 (ImageNet Large-Scale Visual Recognition Challenge). One of their main characteristics is their ability to solve the vanishing gradient problem of deeper networks.

This research paper proposes the use of residual networks, more specifically the ResNet-18 neural network, along with certain effective preprocessing techniques to increase the sample set and improve the model's prediction performance. This research uses log-scaled spectrograms of the participants' speech as input features to ResNet-18. Two other Convolutional Neural Networks (CNN), Base-CNN and AlexNet, were em-

[4]NiteLine: `https://niteline.ie/`

ployed to use as baseline models to compare the performance of the ResNet. AlexNet was employed for its promising use in audio classification.

Research Objective: To investigate the potential of residual networks in detecting if a person is depressed or not based on their PHQ-8 scores while using only their speech as input.

To investigate the above research objective, this study will implement various ResNet models including a standard CNN and AlexNet for comparison. They will be trained on spectrograms of audio recordings, and their effectiveness in prediction will then be evaluated using interpretable metrics such as F1-score, precision, and recall. These metrics will then be compared with state-of-the-art research conducted in this domain using the same DAIC-WOZ dataset. While this study uses the DIAC-WOZ corpus which consists of American participants, it is possible that it might not fair well with other accents of the English language.

The rest of this research paper is organised as follows: Section 2 reviews state-of-the-art research performed in depression detection and provides the rationale for studying residual networks; Section 3 describes the methodology carried out in terms of data acquisition and its preprocessing; Section 4 provides an overview of underlying design architecture of this research; Section 5 describes the implementation of the Base-CNN, AlexNet, and ResNet-18; Section 6 critically analyses the base models and ResNet-18, and provides a detailed discussion of its implications; finally, Section 7 provides a brief conclusion of this research study with viable areas of future work.

2 Related Work

2.1 Depression Detection Methods and Speech

In recent research, we have seen data from different domains be used to aid in depression detection. Traditional machine learning algorithms and, more recently, neural networks have been used on a variety of data sources in attempts to study and predict the likeliness and severity of Major Depressive Disorder (MDD) from early signs of depression. This research has been conducted on datasets of posts from social media and online forums. Facial expressions (Zhu et al.; 2020) have also been used for prediction. Data of participants such as Electroencephalograms (EEG), Electrocardiograms (ECG) (Zhu et al.; 2020), MRI scans (Mousavian et al.; 2019), and other physiological data (Zhang et al.; 2020) have been regularly used in past research in depression detection.

The methods mentioned above use data that require obtrusive means of retrieval. This poses a difficulty in obtaining accurate data from clinically depressed individuals. Speech has recently been widely used in research to predict and analyze signs of depression in individuals. This provides for a more unobtrusive way to detect signs of early depression which could ensure that appropriate attention and help are provided to at-risk individuals. Moreover, it is cost-efficient to record speech than to require individuals to obtain physiological data such as ECGs, EEGs, and MRI scans.

One of the advantages is that the paralinguistic features of speech such as prosody, tone, and pitch make it easier to discriminate individuals diagnosed with clinical depression from individuals who do not suffer from depression. Huang et al. (2020) proposes

a method that analyzes the acoustic features such as prosodic, spectral, and glottal features along with abrupt changes in speech articulation to detect depression in individuals. Such types of feature extraction have been proven to be the norm when using speech for depression detection.

2.2 Use of Spectrograms

Spectrograms are a visual method of representing the amplitude of a signal (loudness) in various frequency spectrums as they vary with time. The amplitude of the signal is represented by the color intensity of the spectrogram. It is usually represented as a 2D matrix of a grayscale image where the values of the matrix represent the amplitude or intensity of the signal.

While prosodic features such as Mel-Frequency Cepstral Coefficients (MFCC), and Zero Crossing Rate (ZCR) have been extensively used in past research we are seeing a growing shift to using spectrograms for audio classification as they contain a high level of detail rather than represent lower-level sound features like MFCCs and other prosodic features. Spectrograms have been gaining traction and are being used increasingly in audio classification. In this project, we will be using log-scaled spectrograms that have been extracted from a widely used signal transformation process called a Short-Time Fourier Transform (STFT) which is a specific type of Fourier transform.

Guzhov et al. (2020) advocate for the use of spectrograms in their research of classifying environmental sounds using the UrbanSound8K dataset. They use pre-trained ResNets on spectrograms of clips of environmental sounds. It is found that using log-powered spectrograms is 10% more accurate than existing state-of-the-art approaches that use MFCCs and also Mel-spectrograms which are those that are Mel-scaled instead of log-scaled. Esmaeili et al. (2018) achieves high sensitivity and specificity by using spectrograms of patient breathing patterns to detect respiratory depression using tracheal sound analysis.

Dinkel et al. (2019) also record the promising results in depression detection when using log-scaled spectrograms while the Mel-scaled spectrograms perform slightly inferior. This could be attributed to the nature of the Mel-scaled spectrograms being less responsive to changes in hyperparameters of the Bidirectional Long Short-Term Memory (BLSTM) model used. Choi et al. (2019) proposes a unique complex-valued spectrogram that follows a different approach of spectrogram extraction from normal straightforward STFT log-scaled spectrograms commonly used in audio classification. They propose to not ignore the phase information when extracting spectrograms from STFT. Magnitude and phase can be accounted for by methods such as phase reconstruction.

Boddapati et al. (2017) also shows spectrograms perform better to MFCCs while using them on popular neural network architectures like AlexNet and GoogLeNet to classify environmental sounds. It is interesting to note that higher accuracy was obtained from spectrograms that were created using a lower sample rate of 8 kHz and unsurprisingly a larger frame length. However, speech signals, unlike environmental sounds, are characterized by frequent abrupt changes in frequency and intuitively larger frame lengths would mean that we lose the ability to analyze these abrupt changes.

2.3 State-of-the-Art Research

In this section, we will focus on state-of-the-art research that has been conducted in depression detection with more emphasis on those which have worked on the Wizard-of-Oz interviews of the Distress Analysis Interview Corpus (DAIC-WOZ) dataset. This dataset consists of 189 audio files of interview sessions between a virtual interviewer, and normal and depressed participants. It will be described in more detail later in this report.

2.3.1 Feature Extraction

The DAIC-WOZ dataset also contains transcripts of the interviews and facial expressions of the participants. However, there will be more focus on features extracted from the participant speech in the papers discussed below.

Yalamanchili et al. (2020), in their research, extract Low-Level Descriptors (LLD) such as spectral and prosodic features from the COVAREP audio processing package. It is dissatisfying to note that no justification is provided for using statistical measures of these features. MFCCs, short-term energy coefficients, and spectral entropy features are extracted from the audio files in Wang et al. (2020). Sentence-level embeddings are preferred over phoneme-level and word-level embeddings in (Haque et al.; 2018). While Yang et al. (2020) uses GANs to augment data to improve depression level predictions.

Log-spectrograms, which are scaled on the log axis, have been used in Vázquez-Romero and Gallardo-Antolín (2020). A spectrogram crop of 4 seconds each at a sampling rate of 16 kHz is used. Spectrograms along with Mel-scale feature bank features are used in DepAudioNet (Ma et al.; 2016) which is one of the most widely used base architectures used for comparison in depression detection from the DAIC-WOZ dataset. Spectrograms have also been used in M.P. et al. (2019) after low-pass Butterworth filters and the Fast Fourier Transform (FFT) is applied to the audio signals. Srimadhur and Lalitha (2020) uses waveforms in their proposed model.

2.3.2 Class Imbalance Solutions

Out of 189 participants in the DAIC-WOZ dataset, 133 are not depressed while 56 are depressed. This posses a class imbalance issue.

An oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE) has been used in Yalamanchili et al. (2020). Here, the minority class (depressed participants) in over-sampled. Random sampling is used in Ma et al. (2016) to address the class imbalance issue. Here, equal random crops of equal length are taken from each participant to minimize person-specific features that might influence the model. From here, an equal number of samples from both classes, depressed and non-depressed, are chosen to be included in the final training set.

2.3.3 Machine Learning Models

Yalamanchili et al. (2020) finds that using SVM coupled with SMOTE analysis provides them an F1-score of 0.74 which performs better than the Logistic Regression and Random Forest algorithms tested. Haque et al. (2018) proposed a C-CNN (Casual-CNN) which provides an F1-score of 0.76 for the combination of audio, visual, and linguistic data of DAIC-WOZ. A 3-layered CNN is proposed in M.P. et al. (2019) which receives spectrograms as input.

Vázquez-Romero and Gallardo-Antolín (2020) uses an ensemble averaging technique to combine the predictions of individual CNNs inspired by Ma et al. (2016) but without the LSTM layer since no improvements were shown. Ma et al. (2016) includes a 1D-CNN stacked behind a Long Short-Term Memory (LSTM) layer and two fully connected layers. In Srimadhur and Lalitha (2020), a spectrogram based CNN is compared with an end-to-end neural network which performs better and receives an F1-score of 0.77 for the depressed class. The authors conclude that as the kernel size increases, the model can learn more nuanced discriminative patterns thereby yielding better performance.

Chlasta et al. (2019) is one of the very few research papers which use ResNets on the DAIC-WOZ dataset and achieve their best performance from a pre-trained ResNet-50 architecture which gave an accuracy of 78% but a poor F1-score of 0.57. This could be attributed to the lack of sampling to address the class imbalance issue when using the DAIC-WOZ dataset.

A Hierarchical Attention Transfer Network (HATN), which comprises mainly of an attention-based encoder-decoder model, is proposed in Zhao et al. (2020) to predict depression levels by using the PHQ-8 scores of participants. A speaker de-identification architecture for depression analysis is proposed in Lopez-Otero and Docio-Fernandez (2020) and compared with a General Adversarial Network (GAN) based approach. A GAN-based architecture called Deep Convolutional General Adversarial Network (DCGAN) is used for data augmentation in Yang et al. (2020) which shows an increase in performance when more augmented data is introduced. Wang et al. (2020) also uses a CNN-GAN based approach called DR AudioNet which predicts the depression levels in the participants. One thing to note is that the Discriminator in the GAN uses the LeakyReLU as its activation function.

2.3.4 Evaluation Metrics

It is found that F1-scores has been used as a measure to accurately evaluate the performance of machine learning models used in this binary classification of depressed individuals. F1-score is the harmonic mean between the precision and recall and these metrics will be used later in this paper. Other metrics such as accuracy perform poorly when it comes to truly describe the performance of the model. It is satisfying to see that most of the important evaluation metrics (F1-score, Precision, and Recall) have been provided in Haque et al. (2018), Vázquez-Romero and Gallardo-Antolín (2020), and Ma et al. (2016). Sensitivity and specificity have also been found to be reliable metrics to judge the models in depression detection research which are present in Haque et al. (2018).

K-Fold cross-validation has been used in Srimadhur and Lalitha (2020) to increase the accuracy of the compared models. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used to evaluate the predicted PHQ-8 scores in Zhao et al. (2020), Yang et al. (2020), and Wang et al. (2020).

Some ICT solutions have also been used to further evaluate the models in real-time. For example, Yalamanchili et al. (2020) uses an android application to assess subjects using the PHQ-8 questionnaire just as found in DAIC-WOZ. This provides an F1-score of 0.74. However, their training model's metrics of f1-score, precision, and recall have been made to appear high by only highlighting those of the non-depressed class. While the credible, yet poor depressed, class metrics is camouflaged at best. M.P. et al. (2019) also provides an ICT solution which involves a python script-powered Raspberry Pi receiving

input from a microphone and uses the model to predict depression from the user's voice recording. One again, we see that only the accuracy of the model is provided with no mention of F1-score.

Table 1: State-of-the-art research using the audio section of DAIC-WOZ.

Author	Model	Precision	Recall	F1-Score
Haque et al. (2018)*	C-CNN	0.71	0.83	0.77
Srimadhur and Lalitha (2020)	End-to-end CNN	0.79	0.77	0.74
Yalamanchili et al. (2020)	SVM	0.7	0.78	0.74
Chlasta et al. (2019)	ResNet-50	0.57	0.57	0.57
Ma et al. (2016)	CNN+LSTM	0.35	1	0.52

*Research utilizes audio, video, and linguistic sections of DAIC-WOZ for prediction.

2.4 AlexNet in Audio Classification

AlexNet (Krizhevsky et al.; 2012) is a Deep Convolutional Neural Network (DCNN) which is believed to be the neural network paved the way for major advancements in the state-of-the-art CNNs we see today. It won the ILSVRC 2012 (ImageNet Large-Scale Visual Recognition Challenge) by beating the runner-up by a significant margin. ILSVRC is something like the Olympics in Computer Vision research. On the ImageNet dataset, it achieved a top-5 error rate of 15.4% with the runner up achieving the next best of 26.2%. AlexNet has recently been applied in audio classification with the increasing use of spectrograms, scalograms, and other image representations of audio features.

In their paper, Jayalakshmy and Sudha (2020), use a pre-trained AlexNet CNN to predict respiratory disorders from scalograms using the Adam optimizer. It uses a standard AlexNet with 5 (convolution + max-pooling) layers followed by 3 fully connected layers. Singh et al. (2019) also uses scalograms of Phonocardiogram (PCG) signals on a pre-trained AlexNet and achieve high-performance results with specificity and sensitivity at 90%.

However, it is important to be wary of the challenges likely to occur in audio classification using AlexNet. Cohen-McFarlane et al. (2020) predict pre-trained AlexNets would perform better on spectrograms which differ from conventional images, however, they make no further attempt to justify this statement. Other effects can be reduced by introducing noise to increase the generalizability of the model.

2.5 ResNets in Audio Classification

Microsoft Research Asia's 152-layered ResNet architecture (He et al.; 2016a) was the winner of ILSVRC 2015 with an incredible error rate of 3.6%. This made them one of the most acclaimed neural networks in computer vision. The intuition behind this deep architecture is that a large number of layers help in learning more complex features. ResNets, with their concept of 'residual blocks', were a solution to the infamous vanishing gradient problem which DCNNs suffered from. This problem was the reason why shallow architectures such as AlexNet and VGG-19 (Simonyan and Zisserman; 2014) were preferred until ResNets came along. ResNets usually come with 18, 32, 50, 101, and

152 layers. They have become popular in audio classification but its use in depression detection has been surprisingly infrequent with the exception of Chlasta et al. (2019).

In Cox et al. (2018), grayscale 2D-spectrograms of radio signals are used to predict signal classes on Search for Extraterrestrial Intelligence (SETI) research data. The research shows that Wide ResNets (WRN) show improved accuracy than DenseNets and their deeper counterparts, ResNet-18 through ResNet-152. This is achieved with 95% lesser parameters than standard deep ResNets. In (Chen et al.; 2019), spectrograms extracted from optimized S-transform (OST) on the audio signals are used in ResNet-50 to predict certain respiratory sounds like wheeze and crackle. A pre-trained ResNet-50 architecture in Le et al. (2019) is used to make use of transfer learning along with an SVM to classify cries of babies and identify those with asphyxia, deafness, hunger, and pain.

3 Methodology

This research will utilize the Knowledge Discovery in Databases (KDD) methodology, initially proposed by Fayyad et al. (1996). This methodology will be the underlying approach to acquire, process, experiment with, evaluate, and ultimately gain insights from data in the domain of depression detection. In summary, this approach entails the data selection, data preprocessing, data transformation, and the application of the proposed neural network and its evaluation.

3.1 Dataset

The dataset used in this research is of the Wizard-of-Oz interviews from the Distress Analysis Interview Corpus (DAIC-WOZ)[5]. It is obtained from the University of Southern California's (USC) Institute for Creative Technologies and was part of the 2016 Audio/Visual Emotional Challenge and Workshop (AVEC 2016) (Valstar et al.; 2016). An End User License Agreement (EULA) was submitted and, upon approval, credentials were provided to access the dataset from which the interview audio files were downloaded. All participants who have taken part in the study in the DAIC-WOZ dataset have signed waivers that approve the usage of the data collected for academic research purposes.

The dataset consists of 189 .wav audio files of interview sessions with depressed and non-depressed participants averaging almost 16 minutes for each interview session. The participant is labelled 'depressed' or 'not-depressed' based on the results of a psychiatric questionnaire, PHQ-8, filled-in by them prior to the interview. A PHQ-score of 10 or more would indicate the participant suffers from a certain level of depression.

These WOZ-styled interviews are conducted by a virtual interviewer named Ellie who is controlled by a human interviewer in another room. The participants are asked open-ended questions such as "Who's someone that's been a positive influence in your life?... Can you tell me about that?".

[5]DAIC-WOZ Dataset: `https://dcapswoz.ict.usc.edu/`

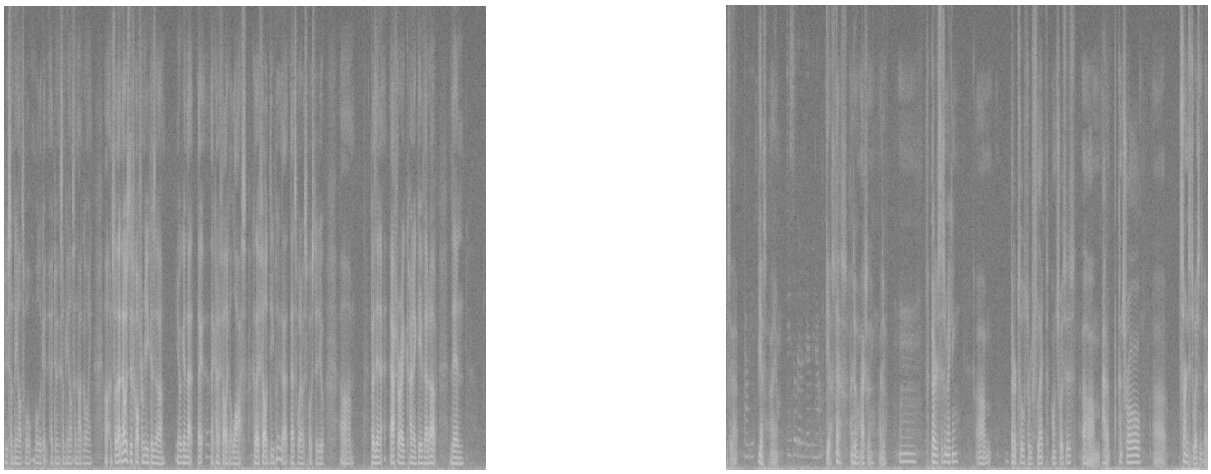
3.2 Audio Segmentation

The files have been segmented to extract only the participant's speech without the silence, background noises, and the voices of other speakers. This was possible since the participants used microphones in a low-noise environment which permitted the segmentation of most audio files with the exception of a few due to technical difficulties. This was done using python's pyAudioAnalysis[6] package.

3.3 Spectrogram Extraction

In this research paper, as mentioned in section 2.2, we have decided to use log-scaled spectrograms of the audio clips. This decision was influenced by the increased efficiency obtained by using spectrograms in recent state-of-the-art audio classification research such as Guzhov et al. (2020), Esmaeili et al. (2018), and Boddapati et al. (2017).

The segmented audio clips, which only contain the participant voices, are then sampled at a 16 kHz sample rate. The Short-Time Fourier transform (STFT) is then performed on the signal using a Hanning window, a frame size of 1024, and a hop size of 512 samples. The resulting spectrogram of each segmented audio clip is obtained in the form of a 2D matrix. They are then scaled logarithmically to produce log-scaled spectrograms. This blog [7] shows that log-scaling of spectrograms perform much better than z-score or min-max scaling. These are then stored locally as grayscale images in preparation for the next preprocessing step, random undersampling.



(a) Participant ID: 482 (Normal)

(b) Participant ID: 426 (Despressed)

Figure 2: 15-second crops of log-scaled spectrograms.

One interesting observation to note is that depressed participants generally provide low-pitched, relatively short, to-the-point, responses to open-ended questions. Whereas, non-depressed participants tend to be more comfortable with providing detailed responses. This behaviour can be visualised in the spectrogram of the depressed participant (b) where their voice signals hardly activate the high-frequency bands unlike with the non-depressed (a) participant.

[6]pyAudioAnalysis: `https://github.com/tyiannak/pyAudioAnalysis`

[7]Blog: `https://medium.com/using-cnn-to-classify-audio/effects-of-spectrogram-preprocessing-for-audio-classification-a551f3da5a46`

3.4 Random Undersampling

Out of the 189 participants in the DAIC-WOZ dataset, a recommend training set of 142 participants and a test set of 47 participants is provided. However, to increase the training set, all participants are merged and random undersampling is performed on the spectrograms obtained in the previous preprocessing step.

Random undersampling is performed for two specific reasons. Firstly, the number of depressed participants is four times smaller than that of the non-depressed participants, and hence the training set and test set must contain an equal number of samples from both classes. Secondly, random undersampling ensures that participant-specific features do not influence the neural network models due to the fact that some interviews are longer than others.

The shortest interview clip spectrogram is chosen and the maximum number of 4-second crops is established. The same number of crops is then sampled from all the other participant interview spectrograms. This resulted in each sample being a spectrogram matrix of 513 rows (frequency bins) and 125 columns (length of 4-seconds) which is the matrix representation of a 513X125 grayscale spectrogram. Finally, the training set is prepared by randomly selecting an equal number of samples (i.e. spectrograms) from both classes, depressed and non-depressed.

3.5 Exclusion of Shorter Interviews

During the experimentation with the three models, Base-CNN, AlexNet, and ResNet-18, it was found that the models were showing poor performance, presumably due to the short training set of 3192 samples. The ResNet-18 architecture obtaining an F1-score of 62%.

This led to increasing the number of spectrogram samples taken from each participant. This was achieved by dropping a percentage of the shortest interviews of both depressed and non-depressed participants to increase the size of the shortest interview. First, 15% of the interviews were dropped, followed by random undersampling, to increase the sample set to 5056 samples. Later, 20% of the interviews were also dropped which increased the sample set to 5520 samples. This set is used as the final dataset to test the models in this paper. The exclusion of shorter interviews, along with the k-fold cross-validation implemented later, is the reason for the increase in the performance of the 3 models.

Excluded Interviews	**No. of Samples**
None	3192
10% of shortest interviews	5056
20% of shortest interviews	5520

Table 2: Spectrogram sample set sizes after excluding shortest interviews.

It should be pointed out that common image data augmentation techniques such as image flipping, shifting, rotation, and zoom would not benefit model performance as a spectrogram is structurally different from object-based images. This means that there is no possibility that there will be a scenario where the model has been inputted a real-time spectrogram that is flipped horizontally or zoomed in.

4 Design Specification

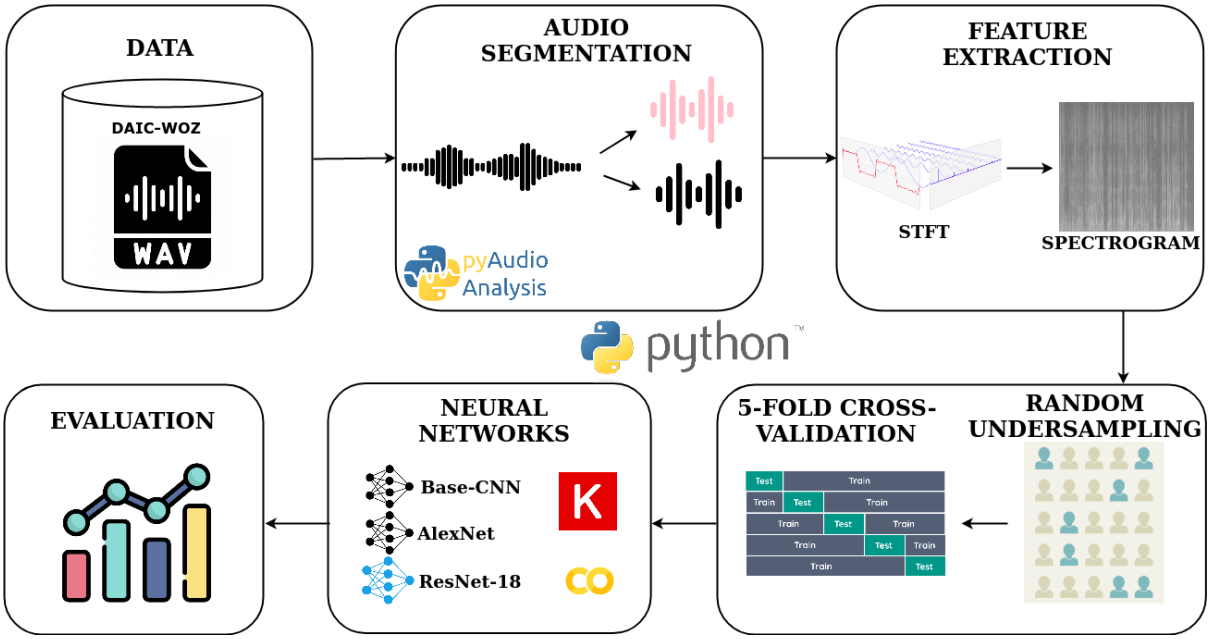


Figure 3: Design architecture of the project.

The above design architecture illustrates the data selection, data preprocessing, model execution, and evaluation of the proposed model using the KDD approach. In Step 1, we acquire the .wav files of the 189 interview sessions. Audio segmentation is then carried out which removes silence, noise, and other speaker voices. The segmented .wav files are then converted to log-scaled spectrograms through a signal transformation known as Short-time Fourier transform (STFT). Random undersampling is then performed on the spectrograms due to the inherent class imbalance in DAIC-WOZ. The 3 models in this research are then trained using 5-fold cross-validation due to the small size of the dataset. The Base-CNN, AlexNet, and ResNet-18 are implemented in Keras, a neural-network library in Python. All steps from data collection to evaluation was carried out in Google Colab. This was due to Colab's provision of a GPU which considerably reduced the training time of the neural networks from hours to minutes. The evaluation includes the critical analysis of the ResNet-18 model advocated for use in depression analysis, and its comparison with other state-of-the-art research carried out using the interview audio files of the DAIC-WOZ dataset.

5 Implementation

The 3 CNNs are implemented in Python using Keras[8] with a tensorflow-backend. Keras is an open-source neural network library in Python. The first, Base-CNN, is a shallow layered CNN. After that, AlexNet and an 18-layered ResNet architecture (ResNet-18) is used to investigate the possibility of improved performance.

[8]Keras: `https://keras.io/`

5.1 Base-CNN

The 2D-CNN applied in this research is used as the base model with which the other DCNNs, AlexNet and RestNet-18, will be compared with. It is inspired based on the paper titled 'Environmental Sound Classification with CNNs' by Piczak (2015). The Base-CNN comprises of 2 (convolution + max-pooling) layers. The output is then flattened and fed to 2 consecutive dense layers, each with 512 neurons. A dropout of 0.5 is added to the output from the 2nd dense layer and is fed to the final output layer which classifies the spectrogram as depressed or not depressed. ReLU activation functions are used in the convolution and dense layers except for the last dense layer which uses the sigmoid activation function for binary classification. Binary cross-entropy is used to calculate the loss function for the Base-CNN, while the Adam optimizer with a learning rate of 0.001 is used.

5.2 AlexNet

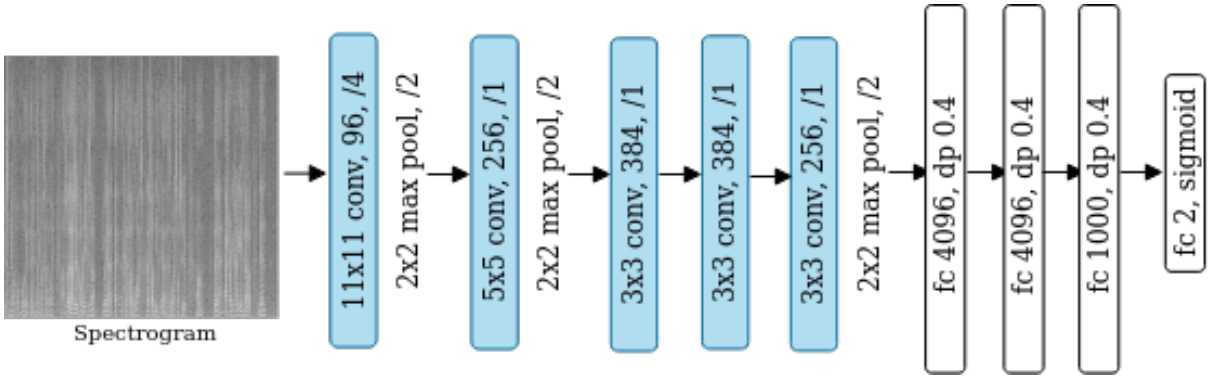


Figure 4: Architecture of AlexNet.

The AlexNet implemented in this research consists of 5 convolutional layers and 3 fully connected layers. Max-pooling is applied to the 1st, 2nd, and 5th convolution layers. The output is then flattened as passed through 3 fully connected layers with a dropout of 0.4. The final output layer contains a sigmoid activation function for binary classification.

Batch Normalization is the the technique of normalizing the output of the activation layer. While it is not usually applied in AlexNet, it was employed in all the layers of this model which increased performance. No padding has been used in any of the layers in this implementation of AlexNet. All convolutional and dense layers use the ReLU activation function while the output layer uses the sigmoid activation function.

The hyperparameters were tweaked and the final version of the model used a 64 batch size, 20 epochs, and Adam optimizer with 0.001 learning rate. Due to the batch normalization, AlexNet receives a lesser training accuracy and therefore suffers less from overfitting than ResNet-18.

5.3 ResNet-18

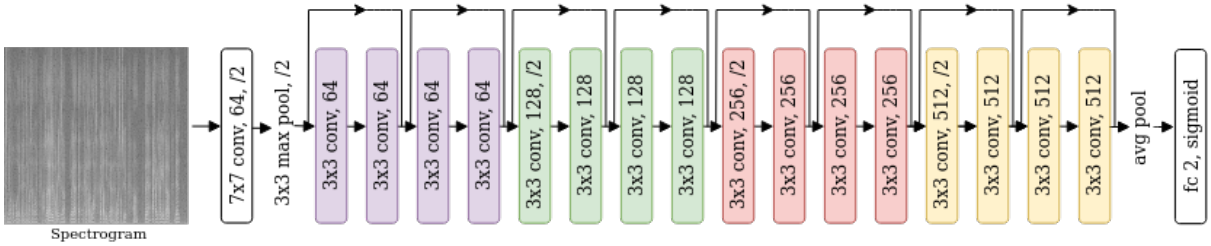


Figure 5: Architecture of ResNet-18 with 'skip connections'.

The ResNet, proposed by He et al. (2016a), was the first Deep-CNN (DCNN) neural network to efficiently tackle the problem of the 'vanishing gradient'. Due to the repeated multiplication of the gradient due to back-propagation, it starts to become extremely small and this leads to the CNN's training loss increasing rapidly after a period of saturation. ResNets use the technique of 'skip connections' or 'identity shortcut connections'. Later, bottleneck residual block (a slight variation of the basic residual block) was introduced by the same authors in He et al. (2016b). A skip connection (see. Figure 5) works by taking the activation from one layer and feed it to another layer while 'skipping' a few layers. This ensures that feature information normally lost or become abstract can still influence the layers later in the network.

The ResNet-18 implemented in this research utilizes the basic residual block instead of the bottleneck variant. Each residual block consists of two 3x3 colvolutional layers, where the input to the 1^{st} layer is added to the output of the 2^{nd} layer before the ReLU activation takes place as implemented in He et al. (2016a). In total ResNet-18 consists of 17 convolutional layers and one fully connected output layer.

The final version of ResNet-18 which provided the best performance based on the F1-score used a 64 batch size, 20 epochs, and an Adam optimizer with a 0.001 learning rate. Higher learning rate values resulted in a decrease in performance due to the model being unable to find the global minima. The model takes an average of 3.5 minutes to train for 20 epochs.

6 Evaluation

6.1 K-Fold Cross-Validation

Due to the inherent class imbalance in the DAIC-WOZ dataset, k-fold cross-validation was employed with a k value of 5. This substantially improved the F1-score of ResNet-18 by more than 9% from 0.75 to 0.83. This is due to the fact that the model is able to learn from sample sets of spectrograms that are better at discriminating between depressed and non-depressed participants. All 3 models were trained using this approach.

It is worth noting that 10-fold cross-validation was also carried out which provided polarizing F1-scores across different folds. This indicates that the test set is too small for 10-fold cross-validation due to its 90:10 ratio of training to test set. Hence, in certain folds, there exists an inadequacy of spectrograms that make the model struggle to predict correctly.

6.2 Base-CNN & AlexNet

Model	Precision	Recall	F1-Score
Base-CNN	0.64	0.66	0.65
AlexNet	0.7	0.8	0.75

Table 3: Evaluation metrics of Base-CNN and AlexNet on DAIC-WOZ.

The Base-CNN that was implemented obtained an F1-score of 0.65. The inadequate performance is to be expected as it was initially designed for detecting environmental sounds (Piczak; 2015) which is structurally different from speech signals.

The AlexNet neural network obtained an F1-score of 0.75. While its performance might not be as efficient as RestNet-18 (F1-score: 0.83), there is reason to believe that shallow networks, subject to appropriate model parameter tweaking, can perform just as reasonably well. Note, that an F1-score of 0.75 means that AlexNet's performance is comparable and even marginally better than Yalamanchili et al. (2020) and Srimadhur and Lalitha (2020) in Table 1.

6.3 ResNet-18

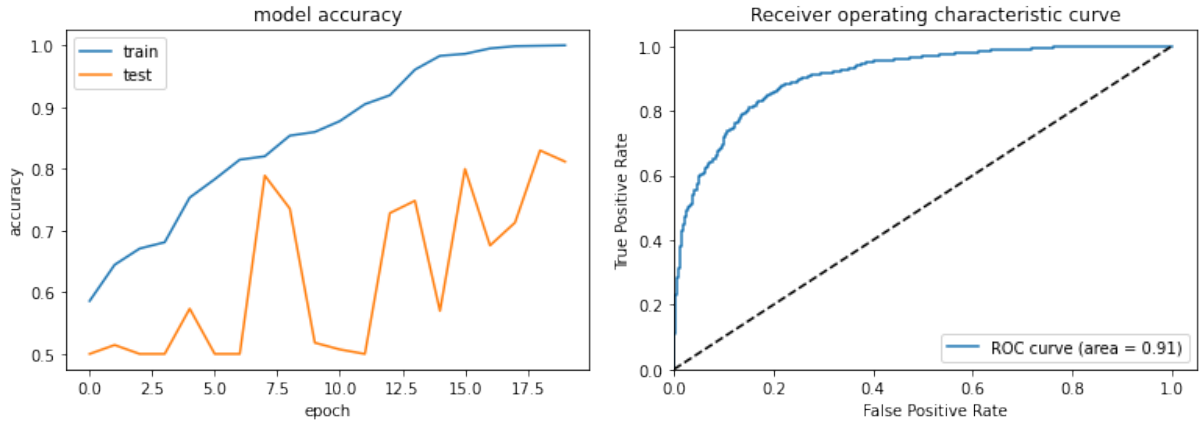
The ResNet-18 model performs significantly better than Base-CNN and AlexNet. This is due to the residual connections that help in amplifying features that would have otherwise become too abstract to be picked up on in the deeper layers. Its 'skip connections' is the reason for its improved performance, and is also the reason why it performs better to most DCNNs which have more than 10 layers.

Model	Precision	Recall	F1-Score
ResNet-18	0.76	0.92	0.83
ResNet-34	0.7	0.88	0.78
ResNet-50	0.73	0.85	0.79
ResNet-101	0.71	0.82	0.76

Table 4: Evaluation metrics of ResNets on DAIC-WOZ.

Other ResNet architectures like ResNet-34,50, and 101 were also implemented with personalised parameter tweaking. Despite the increase in the depth of the architectures, the other ResNets show a decrease in performance. This, coupled with the increase in training time for the deeper ResNets, make them impractical for use in real-time applications such as listening services and emergency helplines where continuous training on new data is required for the model to adapt and for its performance to not deteriorate.

Instead of the standard 4-second 513x125 spectrograms, 16-second crops of 513x513 spectrograms were also created to explore the possibility of improved performance. However, the performance of the 3 models did not show any improvement.



(a) Accuracy (b) ROC Curve

Figure 6: Accuracy and AUC of ResNet-18.

The accuracy plot (Figure 6a) of ResNet-18 shows that the model reaches 100% training accuracy but only gets an 81% testing accuracy. This indicates a case of overfitting which must be addressed in future work. Certain techniques of introducing noise to the spectrograms can be carried out to investigate the possibility of improvement in performance.

From the Receiver Operating Characteristic (ROC) curve (Figure 6b), we can see that the AUC (Area Under the ROC Curve) is 0.91. This is significantly higher than the AUC of 0.85 by Yalamanchili et al. (2020). The high ROC value indicates the model's ability to precisely distinguish between the depressed and non-depressed classes.

6.4 Discussion

In this section, the performance of the ResNet-18 model will be discussed and its implications analyzed. In this research, accuracy will not be used as the principal measure to evaluate the performance of ResNet-18. Precision and recall are far better and more interpretable[9] metrics which indicate how good the model performs in predicting specific classes in relation to others. F1-score, the balance between precision and recall, is the key measure of performance which will be examined. The F1-score of 0.83 for ResNet-18 indicates that the model performs exceptionally better than all the state-of-the-art models discussed in the literature review, as seen in Table 5. It must be noted that appropriate hyperparameter tweaking, along with the usage of 5-fold cross-validation, and the increase of the sample set by excluding short interviews has contributed greatly to the high F1-score by ResNet-18.

The high precision of 0.76, which is only marginally lesser than 0.79 by Srimadhur and Lalitha (2020), indicates the percentage of instances that the model correctly predicted a 'depressed' participant out of all the 'depressed' predictions it has made. That is, 76% of the depressed predictions were correct. The recall of 0.92 indicates the percentage of instances that the model correctly predicted a 'depressed' participant out of all the 'depressed' participants in the entire test set. That is, 92% of the total depressed participants are predicted correctly. The high recall and comparable precision make them

[9]Medium: `https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c`

Table 5: Comparison of ResNet-18 with top-3 models in terms of F1-scores from Table 1.

Author	Model	Precision	Recall	F1-Score
Haque et al. (2018)*	C-CNN	0.71	0.83	0.77
Srimadhur and Lalitha (2020)	End-to-end CNN	0.79	0.77	0.74
Yalamanchili et al. (2020)	SVM	0.7	0.78	0.74
	ResNet-18	**0.76**	**0.92**	**0.83**

*Research utilizes audio, video, and linguistic sections of DAIC-WOZ for prediction.

powerful for use, for example in listening and emergency services. In this scenario, it is crucial to diagnose correctly most of the people with actual depression (i.e. recall), but it would not hurt to have a slight increase in false positives (diagnosing non-depressed people as depressed) which only ensure that more care is taken.

The specificity (percentage of non-depressed people being correctly diagnosed as such) of ResNet-18 is 0.71 is much higher compared to 0.66 by Haque et al. (2018) which also used facial expressions of the participants from DAIC-WOZ. This method of using only spectrograms is non-invasive as no information of gender or age is preserved which increases its applicability in various domains that adhere to stringent regulations of data privacy (e.g. GDPR in Europe).

While the results may seem encouraging, there are some drawbacks of this research, and the data itself, that must be examined in future research. The ResNet-18's high training accuracy of 100% and relatively low testing accuracy of 81%, as seen in Table 6a, indicates that the model is experiencing a serious case of overfitting. This could be due to imperfect preprocessing, for instance, where the voice of the interviewer or other speakers may not get filtered by the audio segmentation performed. This could be remedied by excluding the first few seconds of the clip to decrease the chances of the assistant's voice from seeping into the segmented audio clip.

Another drawback in this research, and more generally with depression detection using only speech, is that people with Major Depressive Disorder (MDD) might sometimes be in a happy mood and their voice might not show signs of depression. This could be the case with the participants in the DAIC-WOZ dataset. It would be interesting to see how this model fairs on those people in their daily life. But then again, the very aim of the application of neural networks, in this research field of depression detection from speech, is to identify those with depression who healthcare professionals, psychiatrists, and counsellors find hard to diagnose. Or at its worst, this model could serve as an indication for the necessity to seek professional help.

7 Conclusion and Future Work

This paper proposed the use of residual networks in detecting if a person is depressed or not using spectrograms of the audio files of interviews. Preprocessing techniques, which included audio segmentation through to increasing sample set sizes, that were employed were a major contribution to the ResNet-18's superior performance. The Base-CNN, and AlexNet achieved an F1-score of 0.65 and 0.75 respectively which proved to be unsatisfactory. The ResNet-18 model, implemented among other ResNet architectures, provided the

best F1-score of 0.83 (see Table 5). With the considerable improvement in performance of ResNet-18 over the existing state-of-the-art models implemented in current research, we can conclude that the research objective has been met and its effectiveness thoroughly evaluated.

This research is to be aimed to be of use in listening services (e.g. NiteLine) and emergency services (e.g. 999 calls) where only the voice of the individual is available. Here, phone conversations can be used to aid in detecting people with depression. This could help service operators provide the necessary care when dealing with depressed individuals. Since the DAIC-WOZ participants use microphones, a much better quality of speech is captured when compared to mobile phones. Hence, this could present some challenges for the proposed model's practical use in listening and emergency services.

The implications of ResNet-18's high F1-score coupled with a high recall provides promise in the rise of the use of ResNets and residual type network architectures for use in depression detection. While the use of residual networks has been explored in environmental sound classification using spectrograms, these environmental sounds are inherently different from speech spectrograms as it is more complex to find patterns in speech signal spectrograms. All the more difficult it is to predict illnesses such as MDD which contain a plethora of nuanced discriminate features. Considering the difficulty in predicting depression from speech alone, the superior performance of the ResNet-18, in relation to the current literature, reinforces the claim for residual networks in depression detection.

There are still many areas in which this research can be improved. This section aims to shed light on those potential areas in which the future work of this research could be carried out. There has been promising research using Generative Adversarial Networks (GAN) for feature augmentation (Yang et al.; 2020)(Esmaeilpour et al.; 2019), and depression severity prediction (Wang et al.; 2020) of audio spectrograms from DAIC-WOZ. ResNets' encouraging potential with transfer learning (Du et al.; 2018) could also be investigated. ResNet-18's performance on more interpretable prosodic features, such as MFCC and ZCR, could be investigated as it is difficult to interpret which spectrogram features are most effective for the neural network. Curriculum learning (Hacohen and Weinshall; 2019), a sampling technique, could also be examined due to its increased popularity in training neural networks.

7.1 Acknowledgements

The author would like to sincerely thank Dr. Muhammad Iqbal, whose expert advice and guidance proved to be vital in the course of this research.

References

Boddapati, V., Petef, A., Rasmusson, J. and Lundberg, L. (2017). Classifying environmental sounds using image recognition networks, *Procedia Computer Science* **112**: 2048 – 2056. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050917316599*

Chen, H., Yuan, X., Pei, Z., Li, M. and Li, J. (2019). Triple-classification of respiratory sounds using optimized s-transform and deep residual networks, *IEEE Access* **7**: 32845–32852.

Chlasta, K., Wołk, K. and Krejtz, I. (2019). Automated speech-based screening of depression using deep convolutional neural networks, *Procedia Computer Science* **164**: 618–628.
URL: *http://dx.doi.org/10.1016/j.procs.2019.12.228*

Choi, W., Kim, M., Chung, J. and Jung, D. L. S. (2019). Investigating deep neural transformations for spectrogram-based musical source separation, *arXiv preprint arXiv:1912.02591* .

Cohen-McFarlane, M., Goubran, R. and Wallace, B. (2020). Challenges with audio classification using image based approaches for health measurement applications, *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–5.

Cox, G., Egly, S., Harp, G. R., Richards, J., Vinodababu, S. and Voien, J. (2018). Classification of simulated radio signals using wide residual networks for use in the search for extra-terrestrial intelligence, *arXiv preprint arXiv:1803.08624* .

Dinkel, H., Zhang, P., Wu, M. and Yu, K. (2019). Depa: Self-supervised audio embedding for depression detection, *arXiv preprint arXiv:1910.13028* .

Du, H., He, Y. and Jin, T. (2018). Transfer learning for human activities classification using micro-doppler spectrograms, *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, pp. 1–3.

Esmaeili, N., Rabbani, H., Makaremi, S., Golabbakhsh, M., Saghaei, M., Parviz, M. and Naghibi, K. (2018). Tracheal sound analysis for automatic detection of respiratory depression in adult patients during cataract surgery under sedation, *Journal of Medical Signals & Sensors* **8**: 140.

Esmaeilpour, M., Cardinal, P. and Koerich, A. L. (2019). Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network, *arXiv preprint arXiv:1904.04221* .

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI magazine* **17**(3): 37–37.

Guzhov, A., Raue, F., Hees, J. and Dengel, A. (2020). Esresnet: Environmental sound classification based on visual domain models, *arXiv preprint arXiv:2004.07301* .

Hacohen, G. and Weinshall, D. (2019). On the power of curriculum learning in training deep networks, *arXiv preprint arXiv:1904.03626* .

Haque, A., Guo, M., Miner, A. S. and Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions, *arXiv preprint arXiv:1811.08592* .

He, K., Zhang, X., Ren, S. and Sun, J. (2016a). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Zhang, X., Ren, S. and Sun, J. (2016b). Identity mappings in deep residual networks, *European conference on computer vision*, Springer, pp. 630–645.

Huang, Z., Epps, J., Joachim, D. and Sethu, V. (2020). Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection, *IEEE Journal of Selected Topics in Signal Processing* **14**(2): 435–448.

Jayalakshmy, S. and Sudha, G. F. (2020). Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks, *Artificial Intelligence in Medicine* **103**: 101809.
URL: *http://www.sciencedirect.com/science/article/pii/S0933365719304981*

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105.

Le, L., Kabir, A. N. M. H., Ji, C., Basodi, S. and Pan, Y. (2019). Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images, *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, pp. 106–110.

Lopez-Otero, P. and Docio-Fernandez, L. (2020). Analysis of gender and identity issues in depression detection on de-identified speech, *Computer Speech & Language* **65**: 101118.
URL: *http://www.sciencedirect.com/science/article/pii/S0885230820300516*

Ma, X., Yang, H., Chen, Q., Huang, D. and Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification, *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, Association for Computing Machinery, New York, NY, USA, p. 35–42.
URL: *https://doi.org/10.1145/2988257.2988267*

Mousavian, M., Chen, J. and Greening, S. (2019). Depression detection using feature extraction and deep learning from smri images, *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1731–1736.

M.P., A., Chander, S., Krishna, B., S.B., A. and Roy, R. (2019). Diagnosing clinical depression from voice: Using signal processing and neural network algorithms to build a mental wellness monitor, *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pp. 1–6.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks, *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, pp. 1–6.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* .

Singh, S. A., Majumder, S. and Mishra, M. (2019). Classification of short unsegmented heart sound based on deep learning, *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–6.

Srimadhur, N. and Lalitha, S. (2020). An end-to-end model for detection and assessment of depression levels using speech, *Procedia Computer Science* **171**: 12 – 21. Third International Conference on Computing and Network Communications (CoCoNet'19). **URL:** *http://www.sciencedirect.com/science/article/pii/S1877050920309662*

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R. and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge, *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pp. 3–10.

Vázquez-Romero, A. and Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks, *Entropy* **22**(6): 688.

Wang, Z., Chen, L., Wang, L. and Diao, G. (2020). Recognition of audio depression based on convolutional neural network and generative antagonism network model, *IEEE Access* **8**: 101181–101191.

Yalamanchili, B., Kota, N. S., Abbaraju, M. S., Nadella, V. S. S. and Alluri, S. V. (2020). Real-time acoustic based depression detection using machine learning techniques, *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–6.

Yang, L., Jiang, D. and Sahli, H. (2020). Feature augmenting networks for improving depression severity estimation from speech signals, *IEEE Access* **8**: 24033–24045.

Zhang, B., Zhou, W., Cai, H., Su, Y., Wang, J., Zhang, Z. and Lei, T. (2020). Ubiquitous depression detection of sleep physiological data by using combination learning and functional networks, *IEEE Access* **8**: 94220–94235.

Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., Tao, J. and Schuller, B. (2020). Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders, *IEEE Journal of Selected Topics in Signal Processing* **14**(2): 423–434.

Zhu, J., Wang, Z., Gong, T., Zeng, S., Li, X., Hu, B., Li, J., Sun, S. and Zhang, L. (2020). An improved classification model for depression detection using eeg and eye tracking data, *IEEE Transactions on NanoBioscience* **19**(3): 527–537.