National College of Ireland

# Integrating audio and text data with deep learning to detect depression

MSc Research Project
Data Analytics

Jacob Benny Packiaraj
Student ID: x22188801

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Jacob Benny Packiaraj | | |
| **Student ID:** | X22188801 | | |
| **Programme:** | Data Analytics | **Year:** | 2023-2024 |
| **Module:** | M Sc Research Project | | |
| **Supervisor:** | Vladimir Milosavljevic | | |
| **Submission Due Date:** | 16/09/2024 | | |
| **Project Title:** | Integrating audio and text data with deep learning to detect depression | | |
| **Word Count:** | 9599 | **Page Count** | 26 |

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:** Jacob Benny Packiaraj

**Date:** 12/08/2024

**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): | |

# Integrating audio and text data with deep learning to detect depression

Jacob Benny Packiaraj
x22188801

**Abstract**

Major depressive disorder is a mental health condition caused due to stress, trauma, biological and medical conditions affecting millions around the world. Timely diagnosis and treatment are crucial in preventing the condition from worsening and helping individuals regain control over their lives. Traditionally, depression was identified through clinical interviews to ascertain his or her mental condition. This study explores a novel approach for automatic depression detection to classify by combing textual and auditory modalities using a meta-learning technique using DAIC-WOZ dataset. The features extracted from transcript data is used in Bi-GRU/ Bi-LSTM and a hyper-parameter tuned CNN model to capture and train audio related features. The prediction from these models is then integrated using meta learner to enhance the classification accuracy. Natural language processing is used to extract the features from transcript file and features like Mel-frequency Cepstral Coefficients, Chroma and Mel-frequency from the audio file is extracted to train the proposed models. Initial result shows that using text features from Bi-GRU/ Bi-LSTM combining with the audio features by CNN has a notable performance improvement compared to the unimodal classification. Features trained in Bidirectional Long short-term memory and the Convolutional neural network can accurately classify majority of the instances with 94% accuracy and the combined model can accurately identify the depressed class with an F1 score of 0.92. This implies that combination of both the auditory and textual information is indeed helpful in the detection of depression as our method takes advantage of these additional sources of information to make the result more accurate and reliable. Potential future work could include the use of other features and modalities to investigate the classification of depressive disorders and integrate the models with cross language and cross-cultural contexts.

# 1 Introduction

Clinical depression is a mental health disorder that can affect anyone irrespective of age, gender and is affecting millions around the world. Person with depressive symptoms can have affected thinking capabilities, dwindling energy, feelings and behaviour. The cause of depression is not limited to a particular trigger but is associated with a person's stressful / traumatic life event, family history, loneliness, illness or injury. From the statistics presented by WHO an estimated 3.8% of the world population undergoes this medical condition impacting 5% of adults and 5.7% of adults above the age of 60[1]. Dublin City University

[1] https://www.who.int/news-room/fact-sheets/detail/depression

published the Physical Activity and Wellbeing (PAWS) study in 2020 to identify the association between physical activity and mental wellbeing among 5500 young adolescents from 79 post-primary schools reported that 4 in 10 who self-reported the depressive symptoms were 47% female and 28% male (Murphy et al., 2020). The spotlight on the burden of depression has intensified since the global pandemic. Covid-19 pandemic and associated lockdowns leads to increased isolation in turn triggers the existing mental health problems. Patients recovered after pandemic reported symptoms of depression due to the toll that it had on the mental health. In 2023, Ireland's Oireachtas health committee has reported that one-in-five patients with long COVID experience anxiety, depression and post-traumatic stress disorder. The latest report from Ireland's Central Statistical Office that conducts survey every 6 years reported 9% of the Irish population aged above 15 experience mild depression[2].

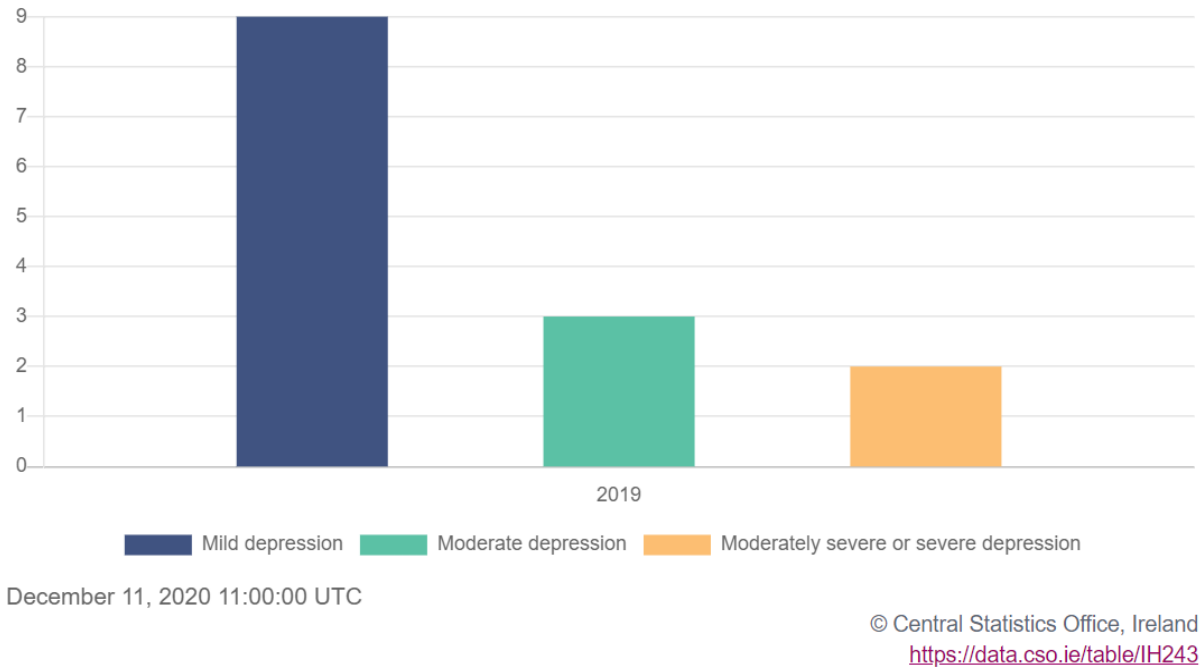


**Figure 1: Percentage by levels of depression reported in Ireland (Age 15 and above).**

Depression, if identified at an initial stage, is treatable, when diagnosed correctly and recovery can begin with a proper treatment plan provided by medical practitioners. Psychological treatment one can receive include therapies like Cognitive Behavioural Therapy (CBT) and Compassion Focused Therapy (CFT) to reduce depression and anxiety and a recent development involves the hybrid approach of integrating the therapy focusing on both behavioural and compassion to reduce the depression and anxiety (Birdsey et al., 2020).

---

[2] https://www.cso.ie/en/releasesandpublications/hubs/p-wbhub/well-beinginformationhub/mentalandphysicalhealth/populationreportingdepression/

However, due to stigma attached with receiving psychological assistance, people tend to ignore diagnosing the condition and leading to life threatening situations. Statistics from European Commission in 2020 stated that 3.7% of all deaths in EU is related to mental and behavioural disorders. A more reliable and accessible tool to diagnose the medical condition can prevent unforeseen situations and applications or models developed using computing technologies can help screen the conditions and provide a support to detect for the presence of depressive symptoms. With huge computational power and rapid advancement in machine learning and deep learning models to train with large corpus of data helps researchers to build applications to classify sentiment or human emotions.

## 1.1 Background and Motivation

Humans tend to express their emotion through verbal and non-verbal cues like eye contact, facial expression, body language, gestures and postures, tone of the voice and speech rate. Other verbal cues include the choice of words used, the intensity of the voice, pauses and silence while holding a conversation about specific topics. These cues can help in identifying potential depressive signs and using this information in automatic assessment can help screen the patient leading to provide immediate support to the affected individuals.

This study explores the classification of depressive symptoms of a person based on processing information from the Distress Analysis Interview Corpus (DAIC-WOZ), a dataset where a clinical interview is conducted on several patients and their response to a questionnaire is recorded and PHQ-8 score is calculated to detect the severity and clinical measure of depressive disorders from clinical studies (Gratch et al., 2014). Wizard of oz interviews is a reliable source of multimodal data where session of the clinical interview is recorded for academic research. Audio, Video and text modality of the interview is stored having rich information of different cues that can be analysed to study the emotional component generated by different participants.

This research is limited to processing transcript and audio files from the participants in this dataset and focus on extracting the features from these modalities to train deep learning models. Natural language Processing will be employed to extract speaker's sentiment and combining with the features from audio that have the speech rate, Mel- frequencies and other component can help better in classifying the depressed and non-depressed state of a person.

How effectively can a GRU/ LSTM model for transcript data be combined with a hyper-parameter tuned CNN for audio data using a meta learning approach to classify depression severity?

This research proposes the use of Bi-directional Long Short-Term Memory (Bi-LSTM) to train the verbal features from the transcript file and a hyper-tuned Convolutional Neural Network (CNN) model to train features extracted from the audio file and late fusion using meta learner to classify depression based on PHQ scores.

# 2 Related Work

Clinical depression is traditionally diagnosed by healthcare professionals by interviewing the patient and by self-reporting mechanism later verified by a valid professional. The current measurement of depression used by practitioners is the Diagnostic and Statistical Manual of Mental Disorders (DSM). The assessment through this method uses a scale to diagnose the forms of mental disorder like suicidal thoughts, mood swings, anxiety, stress, etc leading to provide counselling to the affected individual. Being practiced by professionals, the process is criticized because of recent changes to the methodology and due to introduction of bias by the clinician, as it requires varied skillset to diagnose the conditions. Self-reporting mechanisms includes the Patient Health Questionnaire (PHQ), a criterion based self-assessment tool to respond to several questions to screen a person to diagnose with mental disorders (Kroenke et al., 2009) where a score above the cut point of 10 is considered a person depressive. This includes a medical professional examine the PHQ score based on several parameters and diagnose the patient. Due to stigma associated in approaching assistance for mental health, patients tend not to diagnose the condition and hence this mechanism can provide a better screening of the conditions.

Automatic assessment of diagnosing depression comprises the use of extracting feature/ features from unimodal or multimodal data, preprocessing the data by applying transformations and extracting required features to identify the depression markers. These features are trained using specific machine learning or deep learning model to categorize the pattern from the provided features thereby generating knowledge to predict a patient's depressive state.

## 2.1 Text based depression Assessment

Researchers widely rely on public social media data available in Twitter, Facebook and Reddit to develop machine learning and deep learning techniques to understand the sentiment from the posts and comments posted by users on these websites. Pandey et. al., (2023) in his paper mentioned preprocessing techniques to extract feature from this content include the use of available learning algorithms- Natural language toolkit (NLTK) and other word embedding algorithms like word2vec, Glove and BERT models to transform the verbal cues from the website to corresponding numeric feature vector that can be interpreted by the learning algorithm and make a prediction.

People undergoing stressful situation tend to share their thoughts and feelings through social media sites via posts, images and videos. To hide their identity, depressive people try to communicate through text and hence this helps to uncover insights related to the symptoms of depression. Nadeem et al. (2016) used data from twitter containing 1,253,594 tweets from 1000 user accounts and used bag of words approach to quantify the dataset and created 846,496-dimensional feature space as an input vector. The chosen machine learning technique in this research includes SVM, decision trees, naïve bayes and logistic regression. The average classification accuracy of around 86% and a precision score of 0.82 and

concluded Bag-of-Words was a useful feature set to quantify the emotions and used for classifying major depressive disorder. A similar work conducted by Tejaswini et al. (2024) using twitter tweets of around 6164 records labelled equally to depressive and non-depressive posts is taken from Kaggle. Pre-processing technique for textual information is carried out and fastText word embedding is used to exhibit better representation of texts with semantic information and a better strategy to construct the n-gram that is not part of the corpus. The FCL model discussed in this paper attained an accuracy of 88% with fastText embedding for the Twitter data.

Dinkel et al. (2020) used the DAIC WOZ dataset and used the transcripts as previous researchers have used large user-generated data and as clinical conversations are less investigated. The researcher proposed a bidirectional GRU model that takes the text embeddings and used a variety of pretrained sentence embedding techniques like word2vec, fastText, ELMo and BERT and achieved a cumulative macro F1 score of 0.84 and MAE of 3.48. Symptoms Network Analysis is another technique used by Milintsevich et al. (2023) which addresses a shift from categorical analysis of depression towards a personalised analysis of symptoms profiles. As part of this work, a multi-target hierarchical regression model is used to capture the fine-grain overview of individual symptoms for each participant and the DAIC-WOZ dataset achieved a prediction score MAE of 3.87. The paper addresses the need for shift to multi-targeted regression techniques scoring each symptom individually but ignores the potential of the limitations of this approach.

A comparative study of different pretrained word embedding techniques gives an idea of the choice of specific technique that can be employed to use the word to a vector format data. Reliability of using DAIC WOZ dataset has advantage compared to social media datasets as not all aspects of depression is captured in text and additional non-verbal cues and facial and body language information is required to build a better classifying system. Also, the effective reliability of social media data is unreliable as the source is not verified on medical grounds and cannot be used to employ the models to classify depression. Classification models relying solely on text data has limitations in the depth of analysis of an individual's depressive condition as full spectrum of data is not captured in the posts from social media and this leads to a better source of dataset for a reliable depressive detection system.

## 2.2 Audio based depression assessment

Speech is a critical parameter for classification as it is easy to capture and process. Audio cues include the patients emotional, psychological, and social goodness and on other feature that aids for better classification of the speech. Affected person speaks with a decrease in verbal activity, speech length, decreased speech rate and increase in silent pause in between the conversation. Features extracted from the audio files include the prosodic feature, source features, formant features, spectral and cepstral features.

Yalamanchili et al. (2020) discusses about the development of a machine learning model for real-time depression detection using acoustic features using this dataset. The classifier uses the prosodic, spectral and voice control features extracted from the COVAREP toolkit. The SVM algorithm achieves 93% accuracy in classifying. The research also used 50 real-time data and the DCM model gives an accuracy of 90%. The work also includes an android application to sample the audio and use for the depression classification. The research addresses the class imbalance by using SMOTE to oversample the scarcity in data for training the model. This is an efficient technique to perform the balance of data so that there is no bias in the development of the machine learning model. Sardari et al. (2021) proposed an audio-based depression detection framework using a Convolutional Neural Network-based Autoencoder (CNN AE) to automatically extract relevant features from raw audio data. The study addressed the limitations of hand-crafted feature extraction methods by reducing high-dimensionality issues and overfitting risks. By utilizing deep learning techniques, the proposed model achieved a significant improvement of at least 7% in F-measure of approximately 0.71 for classifying depression compared to existing methods. The framework was evaluated on the DAIC-WOZ dataset, demonstrating its effectiveness in accurately detecting depression. The paper addresses the class imbalance in the dataset using cluster-based sampling technique. The use of deep learning for the classification of audio features is highlighted in this research and includes the benchmarking with the SVM classifier.

DepAudioNet is another novel deep learning model for audio-based depression classification that addresses the challenges in data representation and sample imbalance (Ma et al., 2016) in DAIC-WOZ dataset. The model combines the CNN and LSTM to create a more comprehensive classification model for depression – related vocal characteristics. Random sampling is used for the class imbalance. The DepAudioNet performs by achieving an F1 score of 0.52 and the validation of this model has classified 100% non-depressed cases accurately.

Predominantly the work of Yalamanchili, Sardari and Ma addressed the sample imbalance in the data as the number of records is very low to design a model used for its application in healthcare industry. The varied use of deep learning models is adopted to evaluate the performance of the models and predominantly CNN performs well in the classification of the depressive cases.

## 2.3 Audio-text based depression assessment

Recent research has focused more on the development of depression detection system using multimodal approaches. Shen et al. (2022) introduced the EATD-Corpus, the first public Chinese dataset containing both audio and text data for depression detection. They developed a GRU/BiLSTM-based model that achieved state-of-the-art performance on multiple datasets. It uses the audio and text features from the generated dataset where the GRU model is used for the audio modality and BiLSTM model for the text modality that

generated an F1 score of 0.77 and 0.83 respectively. The multimodal model fusion method outperforms the methods using single modalities, to the prior research conducted.

Another research conducted recently by Iyortsuun et al. (2024) proposed Additive Cross-Modal Attention Network (ACMA) for depression classification using Audio and textual data from the DAIC-WOZ and EATD Corpus. The proposed methodology used Bi-LSTM model with attention for the interview transcripts and use the mel-spectrogram to another Bi-LSTM model with attention, concatenated by ACMA network to classify the depressive categories. SMOTE is used to balance the data for the model training purpose. The model with attention mechanism performed little better compared to the model without the attention. F1 score of 0.8 is achieved with the DAIC-WOZ dataset and 0.78 for the EATD-Corpus.

Solieman and Pustozerov et al. (2021) in their research also used the bimodal analysis of the depression detection where the word level feature is extracted using Natural Language Processing (NLP) techniques and a voice quality analysis model on tense to breathy dimension. The COVAREP data is used from the audio feature and used in a CNN and LSTM model to classify the depressive cases. The concatenated model has an accuracy of 0.68 with the F1 score of 0.79 for the non-depressed class and 0.35 for the depressed class. Similar work is carried out by Lin et al. (2020) where he proposed Bidirectional LSTM for linguistic content, one-dimensional CNN for speech signals and a fully connected network for integrating the output is developed using DAIC WOZ dataset. The audio features like Mel Frequency Cestrum Coefficient (MFCC), COVAREP features, and Mel-spectrogram is used to train the model. ELMo embeddings are taken from the transcript to extract text features. The overall fusion model classifies the task and predict with an F1 score of 0.81 and MAE of 3.75.

Shen et al. and Ngumimi et al. used the EATD-Corpus and DAIC -WOZ where the implementation od deep learning is predominantly focussed and for the dataset under this research got an overall score of 0.85 F1. In the research by Ngumimi, the model with Attention outperforms the model without attention and in overall the prior model by Shen et al. achieved comparatively higher F1 score. The deeper analysis of this model uses resampling techniques to balance the dataset and ELMo embeddings. NetVLAD is used for the corresponding audio to train the model.

In the work on Lin et al. the BiLSTM – 1D CNN model is proposed with audio features of mel-spectrogram is used as it is regarded as a non-linear transformation of spectrograms that has more detailed levels in sound files. They also proposed a new resampling method to balance the samples from two classes. Multimodal fusion of these two models is concatenated horizontally. The results obtained looks promising from this model compared to the other research works conducted by other researchers.

## 2.4 Proposed Solution

Several limitations identified in prior research, such as a focus on single modalities, small datasets, and simplistic feature extraction, were addressed in this thesis.

Earlier works often focused on either text or audio data alone, potentially missing out on the nuanced ways depression manifests through multiple channels. This research adopts a multimodal approach, integrating both text and audio features, allowing for a more comprehensive analysis of depression symptoms. A recurring issue in previous studies was the use of small or homogenous datasets. These models were prone to overfitting and struggled to generalize well to real-world data. This thesis overcomes this limitation using data augmentation techniques such as SMOTE for oversampling the depressed class and NLP-based paraphrasing techniques to increase the diversity of the text data.

Previous methods often relied on basic features, potentially missing complex patterns related to depression. This thesis introduces advanced deep learning models (Bi-GRU/Bi-LSTM for text, CNN for audio) to capture richer and more nuanced features, significantly improving classification performance. The severe imbalance in datasets led many models to overlook depressed cases, reducing their efficacy. The use of synthetic oversampling (SMOTE) effectively balanced the dataset, improving the model's ability to detect depression, especially in minority classes, thereby reducing false negatives.

The literature review demonstrated significant advancement in multimodal diagnosis of depression. Automatic assessment methods leveraging audio and text data has identified the symptoms. However, reliability of these methods can be compromised due to the availability of data and limitation with single modality approaches. Multimodal models combine the audio and text modality has emerged as better solutions but can expect limitation due to data imbalance and complexity of integrating data sources. The data augmentation and balancing techniques to address the imbalance in the dataset can be achieved by using Synthetic oversampling methods to ensure the models are trained on balanced and representative datasets, thereby improving reliability and accuracy if the proposed model. SMOTE is a more reliable method to handle the imbalance

To capture long term dependencies within text, which is crucial for understanding the nuance of language, a choice of bidirectional nature of the deep learning models allow the text to read in both forward and backward directions so by capturing from both past and future states within the sequence. The proposed deep learning approach to capture the feature from the transcript use the Bidirectional GRU/ LSTM model as they are suited for processing sequential data. The study of previous works suggests the use of Recurrent Neural Network models as it is well suited for sequential data and the proposed models should capture the features extracted and make a classification. Features like spectrograms from the audio data can be represented as images with time on the horizontal axis and frequency on the vertical axis. Other features can be included is the Chroma features and the MFCC coefficients which could give added benefit in classifying as the features contain the emotional part of the audio cues which could potentially better understand the audio signal. CNN specifically used for

extracting spatial features from these spectrograms is possible to detect variations related to pitch, tones and potential patterns related to mental health status. It is also useful in distinguishing specific frequencies and high-level features like prosodic features used for depression classification.

The research suggests the use of meta-learner to concatenate the response predictions from the Bi -GRU/ LSTM and CNN models to combine the strengths from both models, leading to comprehensive understanding of the depressive state of the patient. This model suggests the leveraging of complementary information where the transcript data provides the context of what is said, and the audio data offers the cues, tone, pitch and speech patterns. Meta learner can combine the features by integrating both modalities leading to a more accurate assessment of depression. Meta learner model using Logistic Regression and Gradient boosting is proposed as part of this solution.

# 3 Research Methodology

The proposed approach for this research exploring the DAIC-WOZ dataset involves extracting the features from the audio and text modalities. The machine learning methodology used in this research is the Knowledge Discovery in Databases (KDD). This approach focuses on discovering patterns related to psychological distress from the text and audio modalities and the methodology provides a well-organised approach in performing a data mining project. The steps include the selection, preprocessing and transformation of the data followed by using a machine/ deep learning model to identify the pattern and is evaluated using key metrics and the knowledge gained through this process is used to make decisions or implement solutions.

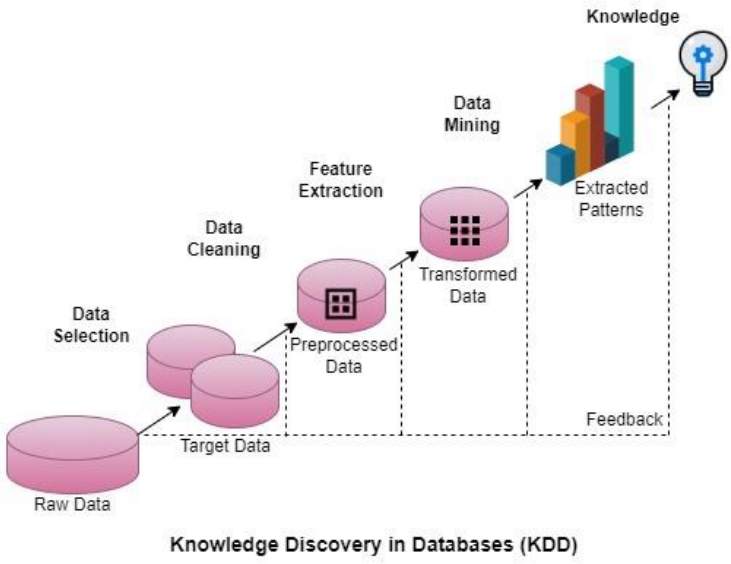


**Figure 2: Knowledge Discovery in Databases process.**

## 3.1 Dataset Understanding

The dataset used in this research is part of the large corpus from the Distress Analysis Interview Corpus (DAIC), a database containing clinical interviews to study and diagnose psychological distress conditions like anxiety, post-traumatic stress and depression. This dataset is a private dataset collected by University of Southern California by the department Institute of Creative Technologies[3]. From the clinical interview conducted, the audio and video recording of the patient is collected and the textual response of the conversation in responding to the PHQ-8 questionnaire is also collected as part of the dataset. Access to the dataset is obtained by submitting an End user license agreement to the university and the credentials were shared to download the dataset and the waivers is signed between the participants and the university to use the data only for academic research. No personal identifiable information is collected from the participants. Dataset contains interview and their response recordings from 189 participants along with the PHQ-8 scores, gender and scores related to sleep, tiredness, appetite and their interest in participating the interview. From the PHQ-8 score levels, participants having a PHQ-8 score of 10 and above is categorized as depressed.

For this research, the textual and audio modalities from the dataset is used and the features were extracted following the KDD methodology to understand the patterns among the participants. The transcript file contains the information about the conversation between the participants and an animated virtual interviewer named Ellie. The csv file contains the conversation between the participants in text format which includes the emotional state of the participant to understand the tone of the response. Similarly, the audio recordings of the participants are used to extract features and train models to classify whether the participant is depressed or not depressed. The audio file for the interview is in .wav format and contains the actual conversation between the interviewer and the participant.

## 3.2 Dataset imbalance

The DAIC-WOZ dataset contains clinical interviews of participants to support and diagnose the psychological distress conditions. The dataset is available to academic researchers as part of the Audio/ Visual Emotional Challenge (AVEC) to provide a benchmark test set for multimodal analysis aimed for the comparison of machine learning models for automatic audio, video and audio-visual health and sensing the emotions with participants in the clinical interview strictly under the same condition. The dataset contains recording and transcripts of participants that undergone clinical interviews with a virtual computer agent. As part of this dataset, a binary level of the score associated with PHQ-8 is provided that represent the presence of depression for each record.

The overall dataset contains 189 records and for the research purpose, it is split into 107 participants for the training set and 35 participants. Out of the records available, 77

[3] https://dcapswoz.ict.usc.edu/

participants in the training dataset are non-depressed and 30 are labelled depressed. Similarly, in the training set, 23 are not depressed and 12 are marked depressed. To understand the performance of the proposed model, the data is split as per the AVEC2017 challenge (107,35) to benchmark with the results from models proposed in other scholarly research works. Few of the techniques researched to counter the imbalance in the dataset is used to resolve this problem.

## 3.3 Preprocessing of Text and Audio files

Transcript file contains the actual conversation between then the interviewer Ellie and the participants and has the values in tab delimited format containing start_time, stop_time and speakervalue. For the proposed research model, the speakervalue is of prime importance as it contains the response of the clinical interview that includes the emotions associated with in the response like "<laughter>" to understand the tone of the response. It is essential to capture the tone as it contains an intrinsic value of the speaker to the text. Transcripts from the dataset is pre-processed by identifying only the response from the participants in the transcript file and ignore the questions by the interviewer. The stop words were removed, punctuation was removed, and most frequent and rare words were removed from the transcript dataset. The word in the conversation is lemmatized, the process of reducing the words to the base or root form. It is one of the important feature engineering techniques in text analysis and natural language processing. Further, synonyms for the word are obtained from the natural language toolkit corpus. The words are then tokenized by the process of tokenization to individual tokens to have a better representation of the words.

The audio file contains the actual conversation during the clinical interview between Ellie and the participant. The file is pre-processed by segmenting it based on the conversation part and the noise and silence part. From the audio file, silence is removed, and the remaining segments is concatenated to create an audio with only the conversations. pyAudioAnalysis library is used to remove the silence and the parameters weight and smoothing are used to adjust the detection of silence from the segmented dataset, identified and is removed from the audio file.

## 3.4 Feature Extraction

This section contains the information on the different features that are extracted from the different modalities that can contribute to the effective classification of depression.

The proposed model uses the technique of word embedding in text analysis, where the technique GloVe is used to extract the features. GloVe – Global vectors another word embedding technique learns from the word vectors by factorizing a word co-occurrence matrix. Both models are utilized in this research to capture the semantic meaning and relationship between the words and is used to pretrain the transcript information. An embedding matrix is constructed to fill it with the features for use in a neural network model as an input.

For feature extraction from the audio file, focus on the extraction of acoustic and prosodic markers from the patient's speech segment. The key acoustic features focussed on this research contain the spectral and cepstral features from the audio file. Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram frequency and Chroma features is calculated from the audio file is collected as the input to the model. An exploratory analysis of the frequency from the Mel-spectrogram of the audio file from the dataset is used. From the segmented audio clips, the silence removed from the audio file is sampled at 16KHz sample rate. In this research, Mel-frequency spectrogram is extracted using the Librosa library. The logic computes the frequency spectrogram, which represent the short-time power spectrum of the sound. This is calculated using a Hann window and is averaged across time to generate a fixed-size feature vector, that captures the overall characteristics of the audio file. This resulting vector is used for the further analysis in the deep learning model.

## 3.5 Justification for Selected Techniques

This research employs the GloVe word embedding technique for textual data, capturing the semantic relationships between words, which is critical for detecting the context-sensitive nature of language used by individuals with depression. The selected features for audio data—MFCCs (Mel-Frequency Cepstral Coefficients), Chroma, and Mel-spectrogram—are well-suited for capturing subtle acoustic properties, such as changes in pitch and tone, which are indicative of depressive states.

**Text Data (GRU/LSTM Models):** The decision to use Bi-GRU and Bi-LSTM models for the transcript data stems from their ability to handle sequential data while maintaining long-term dependencies. Depression often manifests through complex patterns in speech, which these models are well-equipped to capture. The bidirectional nature of these models ensures that information from both past and future contexts is preserved, making them effective in analyzing long conversational sequences typical of clinical interviews.

**Audio Data (CNN Models):** CNNs were chosen for their ability to handle spatial hierarchies in data, making them ideal for processing time-frequency representations of audio signals, such as Mel-spectrograms. The CNN model captures subtle features of speech, including pitch, tone, and rhythm, allowing it to detect depressive speech patterns with high accuracy.

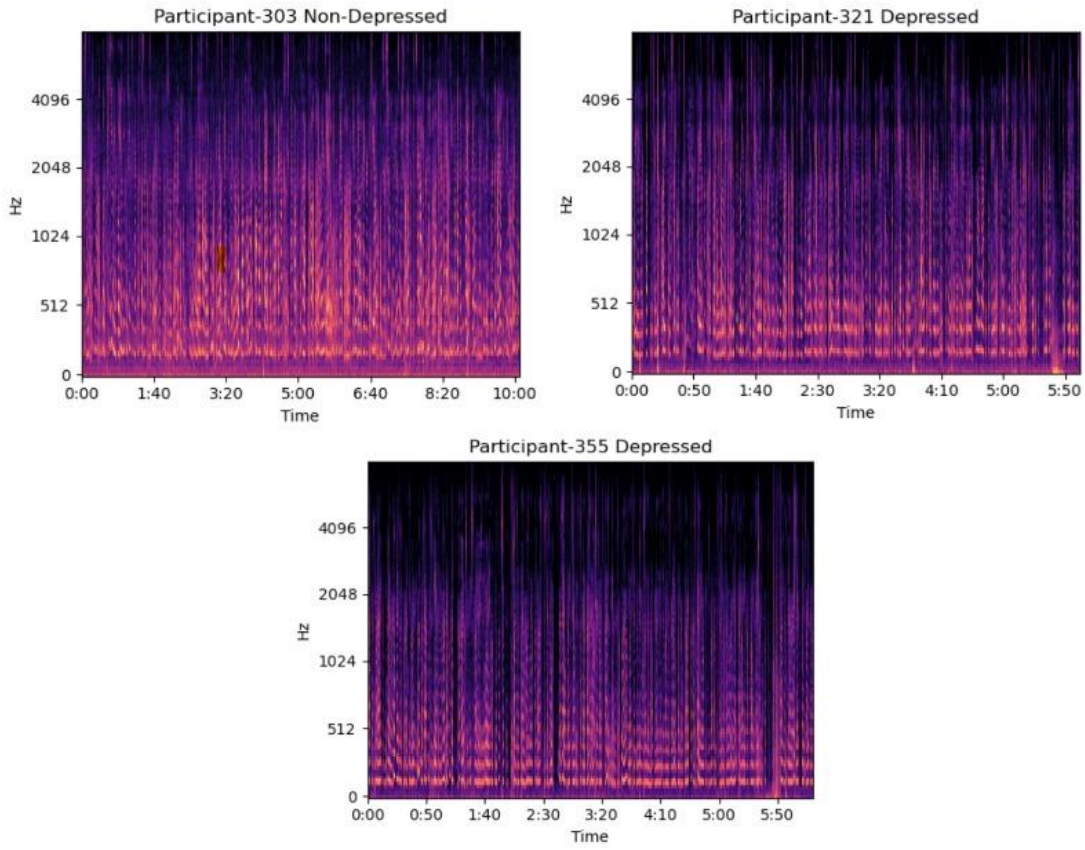


**Figure 3: Mel-frequency spectrogram from the audio files.**

## 3.6 Modelling and Implementation

The extracted features are split into the input and target variables and fed into the proposed deep learning models. The transcript data is used in the proposed bidirectional GRU or in the bidirectional LSTM model to train and the predictions are validated using specific evaluation metrics used for the classification. Similarly, the audio features obtained from the audio file is used as the input to the hyper-parameter tuned CNN model. The layers, dropout rate and learning rate are fine tuned to obtain a model with highest accuracy to be used for prediction of the testing dataset.

## 3.7 Evaluation

The performance of the model is evaluated by an empirical process that is used for the classification by Accuracy (Acc), Recall (Rec), Precision (Pre). Further performance metrics include the F1 score and Confusion Matrix. True Positive (TP) and True Negative (TN) is the number of instances where the model predicts a participant having depression and non-depression state accurately. False Positive (FP) and False Negative (FN) is the “false alarm” and “miss” where the model incorrectly predicts a participant having the depressed/non-depressed state.

# 4 Design Specification

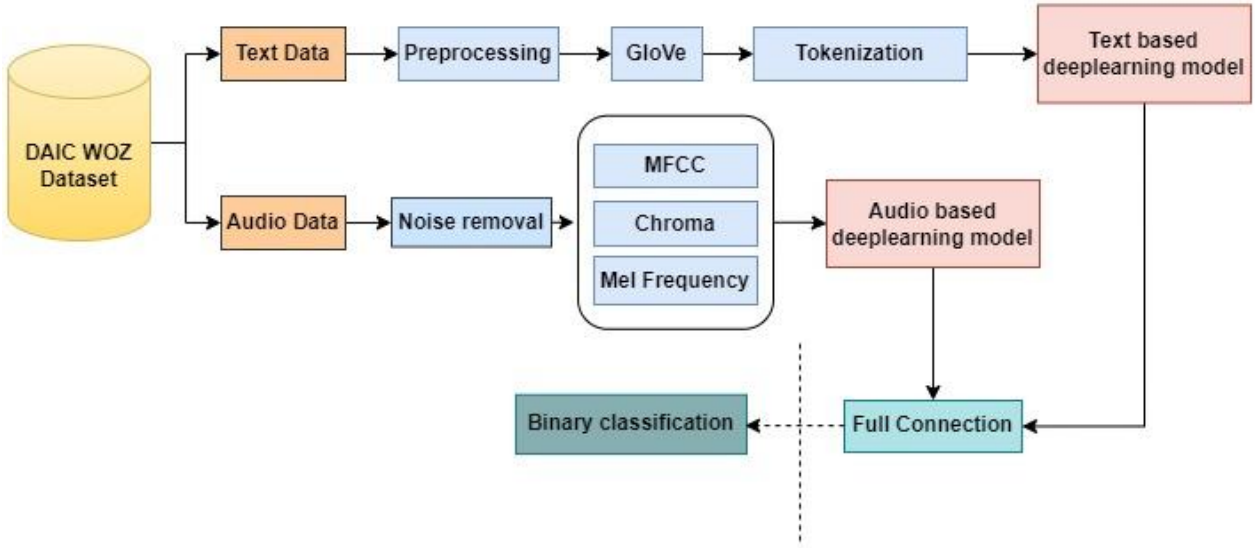


**Figure 4: Architecture diagram for the proposed Meta learner classification model.**

This section includes the methods employed in the study of depression classification using deep learning. Following the KDD approach, in the design architecture, the data is extracted from the source dataset and is selected, pre-processed, models designed and executed and finally the evaluation of the proposed models are verified.

In the text modality, the transcript information is pre-processed by focusing only on the participant's response and whereas punctuations, stop-words and sparingly used words are removed from the dataset. The words are replaced with the synonyms and is then tokenized to be used in the GloVe to obtain the vector representation of the words. The deep-learning models identified for processing this text modality is the Bi-directional Gated Recurrent Unit (bi-GRU) and Bi-directional Long Short-term Memory (Bi-LSTM). 5-fold cross validation is implemented in both the models to train the model with the limited dataset and SMOTE is applied to generate synthetic samples to address the imbalance in the dataset.

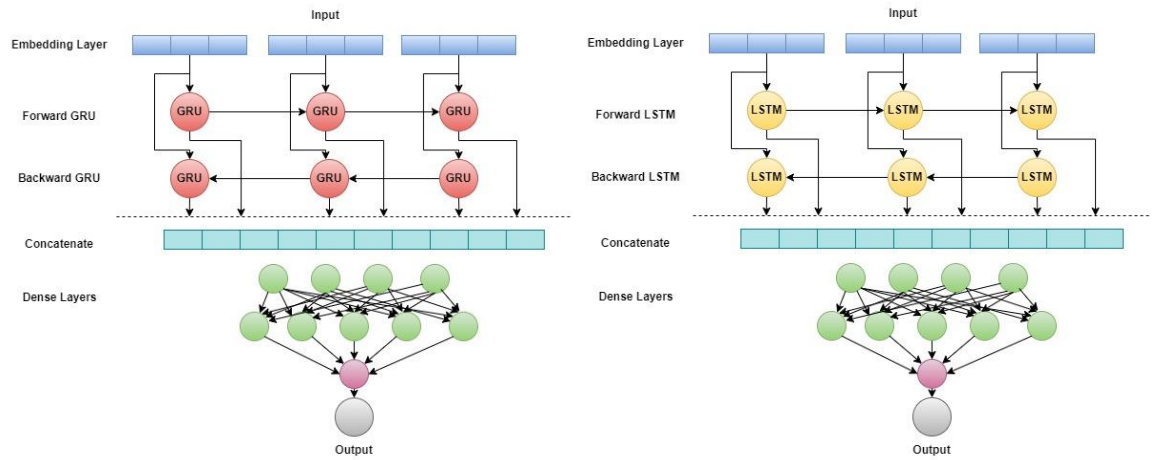


**Figure 5: Architecture of Bidirectional GRU and Bidirectional LSTM model.**

From the audio modality, the data in the audio file is segmented and the silence is removed to capture only the context from the dataset. Mel-frequency Cepstral coefficient, Chroma features and Mel-frequency spectrogram is extracted as features to be used in the classification for the proposed deep-learning model. A custom CNN model is implemented with three branches corresponding to the different audio features. Keras tuner's 'RandomSearch' is used to optimize the model's hyperparameters, to optimize the validation accuracy.

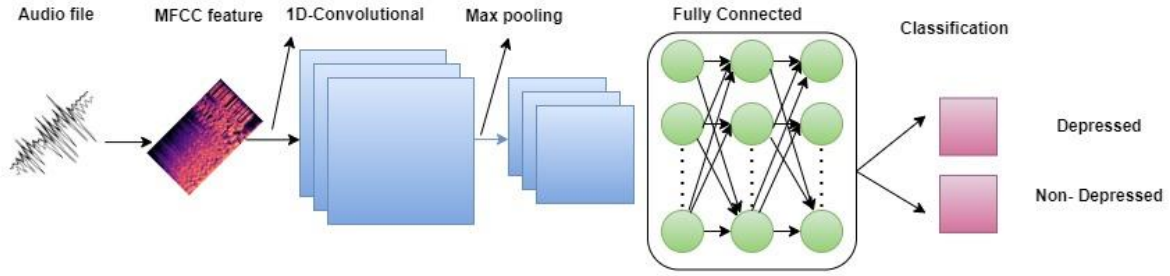


**Figure 6: Architecture of a CNN model.**

# 5 Implementation

This section contains the detailed implementation of the preprocessing steps carried out and the deep learning models used for the text and audio modality to classify the depressive conditions. For the classification from transcript, Bidirectional Gated Recurrent Unit (Bi-GRU) and Bi-directional Long Short-term Memory (Bi-LSTM) is modelled. For processing the audio file, Convolutional Neural Network (CNN) is used. To combine the prediction from the deep learning model between the modalities, a custom meta-transformer using logistic regression is built to classify the depressive nature for the records in the test dataset.

## 5.1 Pre-Processing

The transcript and Audio files are pre-processed to extract the features and train the model. The dataset is split into the train and test data-frames based on the participants listed in the files available in train_split_Depression_AVEC2017 and dev_split_Depression_AVEC2017. The conversation from the transcript file is loaded and read using Pandas library and only the participant's response is captured to extract the features. Initial preprocessing steps are carried on the dataset and the punctuations, stop words and the most used/ less used words are removed. Spacy is used to load the pipeline en_core_web_sm to lemmatize the data by converting the words to it root form and each word is tokenized for a better representation. wordnet library from natural language toolkit is used to get the synonyms of the words and is replaced in the training dataset. The vector representation of the words is obtained from the GloVe and the training data is pretrained to generate the embedding matrix with the words represented in vector format. This data is used to train the proposed deep learning models for the textual data classification.

The raw audio file contains the conversation between the participant and the interviewer. The audio file is segmented using pyAudioAnalysis module and the silence part in the audio is captured and removed from the data. The segments are again concatenated and converted to .wav format using wave module. Features extracted from the audio file includes the MFCC coefficients where 40 coefficients from the audio file are captured using the Librosa library. Similarly, Chroma features is extracted from each individual file where the short-time Fourier transform (STFT) is calculated on the audio signal which provides the time-frequency representation of the audio signal. 12 chroma bins are chosen for this feature extraction corresponding to the 12 semitones of the musical octave. The result is transposed to 12-dimensional feature vector. Chroma features are effectively used for speech emotion detection. It captures the harmonic content and can be sensitive to pitch class thereby making them valuable in recognizing emotional tones form speech.

Rather than using the Mel-spectrogram images from the audio file for classification, in this approach, the features of the Mel-frequency spectrogram is extracted. The array generated is the representation of the intensity of the audio signal across different Mel frequency bands over time. This feature helps in the process of recognising and classifying the audio features for emotion detection.

## 5.2 Bi-directional Gated Recurrent Unit

The first deep learning model chosen for training and evaluating the pre-processed transcript data is bi-directional Gated Recurrent Unit (Bi-GRU). Various modules from Keras package are used to build the deep learning model. For the class imbalance in dataset, Synthetic Minority Oversampling Technique (SMOTE) from Keras's 'imblearn' is used to generate synthetic samples to balance the dataset during the training of the model. From Scikit-learn, 5-fold cross validation is integrated to the existing model to split the dataset into 5 folds with shuffling enabled to ensure randomness while training the model.

The model consists of an embedding layer, GRU layers and GlobalMaxPooling layer, dense and dropout layer and an output layer. A pretrained weights from 'embedding_matrix' is used in the embedding layer where the data from the training dataset word indexes is converted into dense vectors. Three GRU layers are added to the model each with 128 units. The output from the first layer is returned in sequences, leading to other subsequent layers to work on the sequence data. After each GRU layer, dropout is applied for regularization. 'tanh' activation function is used in the subsequent GRU layers. GlobalMaxPooling1D is added to downsample the input representation by taking the maximum value along three dimensions, thereby reducing the dimensionality. A dense layer with 256 units and ReLU activation is used and finally the final dense layer with a single unit with sigmoid function is used for the binary classification.

Callbacks attributes like 'ReduceLROnPlateau' and 'EarlyStopping' are defined in the model designing to adjust the learning rate and to prevent over fitting. Adam optimizer is used to compile the model and a learning rate of 0.004 and binary cross entropy loss is used

in the model. The model is trained on the sampled data with a batch size of 10 for up to 20 epochs. The performance of the trained model is evaluated on the testing dataset.

## 5.3 Bi-directional Long Short-term Memory

Another deep learning model to process the transcript records for the classification is through the Bidirectional long short-term memory (Bi-LSTM). A model like the proposed GRU model is designed to receive the inputs from the training and testing dataset, except, this model uses the Bidirectional long short-term memory is implemented. 5-fold cross validation is implemented to the model to train the model in splitting the dataset with shuffling enabled to ensure randomness. The model is designed with an embedding layer with the pretrained weights, followed by the bidirectional LSTM models with 128 units each with tanh activation function. Each LSTM layer is followed by dropout layers and globalMaxPooling is used for the dimensionality reduction of the features in the model. Finally, a dense layer with sigmoid activation function is used to generate the binary classification of the model. The compilation of the model requires the Adam optimizer with a learning rate of 0.004. Callbacks are made to optimize the training and performance metrics is recorded for each of the folds during the training of the model.

## 5.4 Convolutional Neural Network

MFCC, Chroma and Mel-spectrogram features extracted from the audio file is trained using the CNN model for the classification is converted into suitable representation of arrays using Numpy module. Numpy and Pandas module are used to manipulate the data to convert into the numpy arrays. The features extracted originally are in 2 dimensions and the CNN models require a 3-dimensional input. Before feeding into the model, SMOTE is performed to nullify the imbalance in the dataset. SMOTE requires a 2-dimensional array where each row corresponds to a sample and each column corresponds to a feature. All the features are flattened to the SMOTE process. After the SMOTE is performed, the data is reshaped back to the original format that is required by the CNN model.

The overall CNN model is implemented which branches for each input type, consisting of convolutional layers, max-pooling, and global average pooling layers. A convolutional layer with tuneable filters from 32 to 128 is modelled, having a kernel size of 3 to 5 with ReLU activation function. To reduce the dimensionality by downsampling, Maxpooling layer is added with a pool size of 2. To concatenate the values from these layers, Global average pooling layer is used to convert into a single vector by averaging the values across the time dimension. The outputs of the three branches are concatenated into a single tensor, by combing the features of MFCC, Chroma and Mel-spectrogram. A dense layer with ReLU activation function takes the concatenated input from the previous step and it contains the dense layer unit of 64 to 256 in steps of 64 for tuning. Dropout layer is added to the model to prevent overfitting by randomly selecting a fraction of input units during training and is tunned between 0.3 and 0.7 with steps of 0.1. Finally, the output layer is added to the model having a dense layer with 1 unit having a sigmoid activation function for binary

classification. The overall model is compiled using Adam optimizer with a tuneable learning rate and have a binary cross-entropy function for the binary classification of the model. The model also has a hyperparameter tuner from keras_tuner module to find the best model during the training of the model.

### 5.5 Model stacking with Meta-Learner

The model includes a stacking ensemble approach where the prediction from the base models is combined and used as an input to a meta-learner model. The predictions from the transcript and audio models are flattened to a 1-dimensional array ensuring they have a uniform shape. The predictions from the models are combined into a 2-dimensional array where each row represents the prediction from both the models for a single sample. The resulting array serves as the feature for the meta-learner.

The combined predictions are split into training and for the meta-learner. A gradient boosting classifier is used as a meta-learner, where number of trees of 100 for the number of boosting stages to be run. Maximum depth of individual trees of 3 relating to the model complexity. The meta learner predicts the entire set of combined predictions. This model provides a more comprehensive assessment of the overall model's performance across the entire dataset.

## 6 Evaluation

The proposed models are evaluated based on the Confusion Matrix and ROC AUC curve for each of the implemented models. The need for the additional model to process the audio and transcript features is evaluated based on the metrics.

### 6.1 Bi-GRU Model

The model is trained with the transcript dataset and the Gated Recurrent unit model having 5-fold cross validation splits the training records into 5 samples and trained. Accuracy and Validation loss is chosen for the metrics to choose the model to evaluate the test dataset in further validated with this model. After training the model, the best model identified is used to predict the test dataset. The confusion matrix of this model is in figure 7.

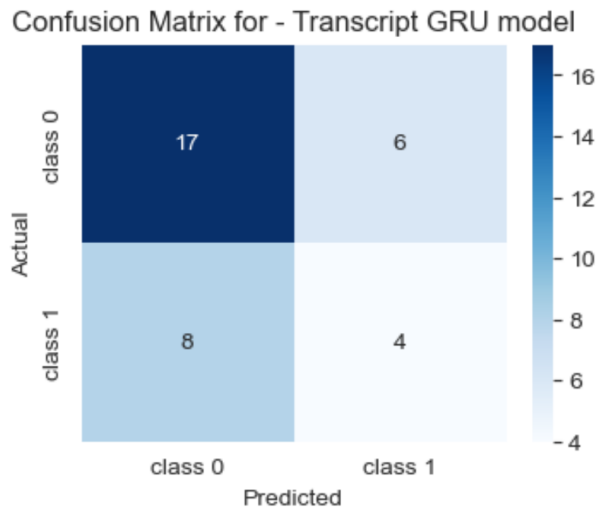


**Figure 7: Architecture of a CNN model.**

17 non-depressed instances were correctly predicted as non-depressed and 4 depressed instances were corrected predicted as depressed. Considering the false scenario, 6 non-depressed instances were incorrectly predicted as depressed and 8 depressed instances were incorrectly predicted as non-depressed. In this model, the true positive of the prediction is not up to the expectation where the precision of the model is 0.40, where 40% of instances predicted as depressed were actually depressed. A recall of 0.33, where 33% of actual depressed instances were correctly identified. The balance between the precision and recall obtained for the depressed class is through F1 score of 0.36. The model struggles in predicting depressed classes which is found in the lower precision, recall and F1 scores of the depressed class. The overall accuracy of the model is 0.66. Another Bi-LSTM model for text classification is identified in this research.

## 6.2 Bi-LSTM Model

Because of the presence of complex patterns in the transcript dataset and to make an informed decision on the models chosen for processing the text data file, bidirectional long short-term memory model is modelled.

The performance of the model is evaluated with the confusion matrix from Figure 8 where true negative of 19, meaning, 19 non-depressed instances were correctly predicted as non-depressed and a true positive of 4, where 4 instances of depressed cases were correctly identified as depressed.

The overall accuracy of the model is around 0.66 with a precision of 0.5 where 50% of the instances predicted as depressed were actually depressed and a recall of 0.33, where 33% of actual depressed instances were correctly identified.

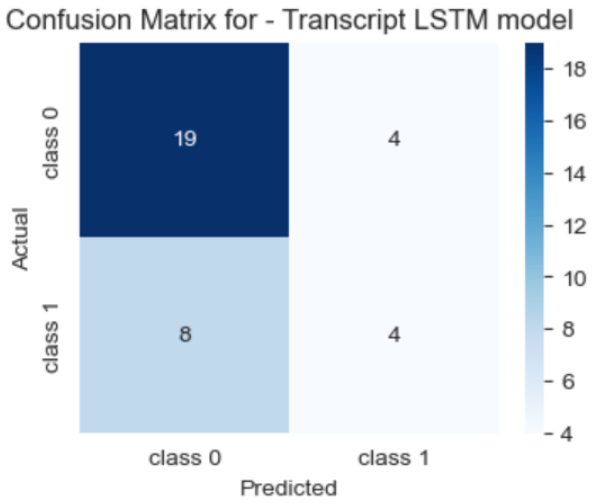


**Figure 8: Architecture of a CNN model.**

The F1 score of the depressed class is 0.40. For the binary classification the key performance metrics is the Receiver Operating Characteristic (ROC) curve. It provides a comprehensive measure of the ability of the model to distinguish between positive and negative classes. The LSTM model has an AUC of 0.73

## 6.3 CNN Hyperparameter model

This model is chosen for processing the pre-processed audio files. The features of MFCC, Mel-spectrogram and chroma are extracted. These features are fed into three individual filters respectively and are classified using the CNN. The basic CNN model is used for the classification of the depressed and non-depressed state of the participants. With the filter count of 64 in the convolutional 1-dimension layer for each of the feature input, with an epoch of 30 for the training.

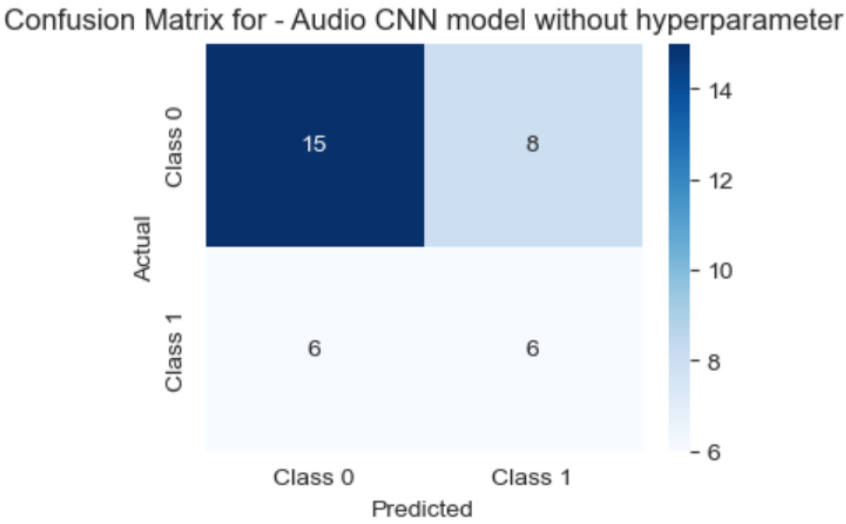


**Figure 9: Architecture of a CNN model.**

From figure 9, the model has predicted 15 non-depressed cases and 6 depressed cases accurately and the false cases is considerably more and has a overall accuracy of 60% with an F1 score of 0.46 for the depressed class.

The final model for the CNN is obtained tuning the hyperparameters from the model with fixed filters and kernel size. The dense layer, dropout rate and learning rate are tuned to get the best model for the CNN classification. There is a trade-off between the true positive and false positive as the tuned parameters model has classified the positive cases more accurately.

There is a significant improvement from the base model to the tuned hyper model with a total accuracy of 69% with an increased precision to 0.56, where 56% of instances predicted as depressed were actually identified as depressed and F1 score of 0.48 for the depressed class.

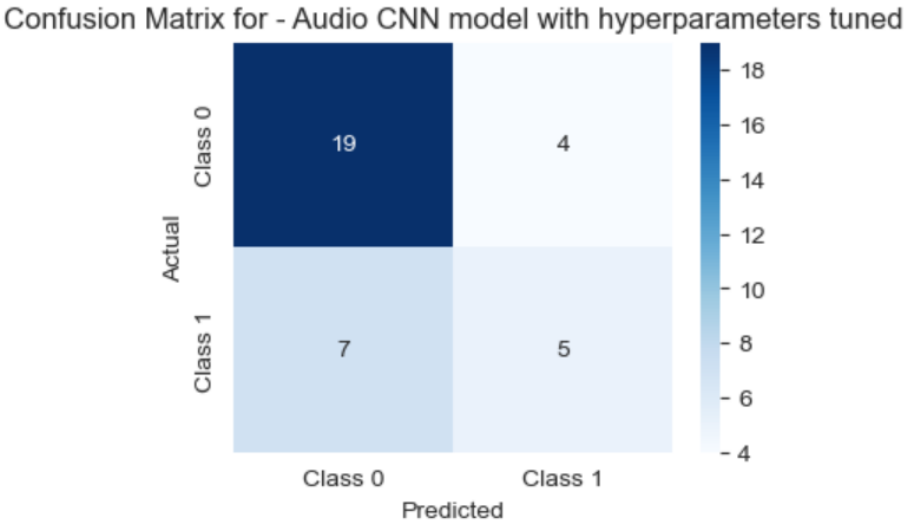


**Figure 10: Architecture of a CNN model.**

Hyperparameters are tuned to identify the best model and is further used for the prediction of the test dataset.

**Table 1: Model Accuracy and metrics**

| Model | Accuracy | Class | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Bi- GRU | 0.60 | Non-depressed | 0.68 | 0.74 | 0.71 |
| | | Depressed | 0.40 | 0.33 | 0.36 |
| Bi- LSTM | 0.66 | Non-depressed | 0.70 | 0.83 | 0.76 |
| | | Depressed | 0.50 | 0.33 | 0.40 |
| CNN | 0.6 | Non-depressed | 0.71 | 0.65 | 0.68 |
| | | Depressed | 0.43 | 0.50 | 0.46 |
| CNN-tuned | 0.69 | Non-depressed | 0.73 | 0.83 | 0.78 |
| | | Depressed | 0.56 | 0.42 | 0.48 |
| GradBoost Meta learner | 0.94 | Non-depressed | 1.0 | 0.91 | 0.95 |
| | | Depressed | 0.86 | 1.0 | 0.92 |

## 6.4 Meta learner concatenation

To combine the prediction from the Audio and text modality, A meta learner classifier concatenating the probabilities of the predictions from the LSTM model for the text modality and the best hyperparameter CNN model after it has been tuned is used for the audio modality. The individual models are trained, and the final predictions are concatenated using the Meta learner where Gradient Boosting classifier is used to train and make the final predictions.

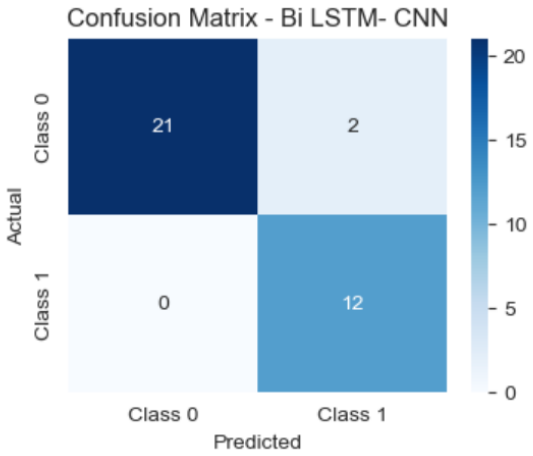


**Figure 11: Architecture of a CNN model.**

The confusion matrix of this model has an overall accuracy of 94% accuracy. In the depressed class classification, 12 instances of the depressed class is identified correctly with a precision of 0.86 and a recall of 100% achieving a overall F1 score of 0.92.

## 6.5 Discussion

To address the imbalance in the DIAC-WOZ dataset, various techniques were explored. Initial efforts focused on class weighting, but this approach proved ineffective, as the model consistently predicted probabilities between 0.1 and 0.2 for the depressed class. Manual adjustments to the weights demonstrated that this method was too simplistic for handling the complexity of the data. Further experiments with SMOTE (Synthetic Minority Over-sampling Technique) yielded better results, significantly improving the classification performance. By generating synthetic samples for the minority class (depressed), the model was able to learn more representative patterns. The use of 5-fold cross-validation helped ensure balanced training and robustness of the results.

In addition to SMOTE, advanced resampling techniques like random oversampling and undersampling were considered. However, oversampling risks overfitting due to repetitive sampling of the minority class, while undersampling may lead to the loss of important information from the majority class. A more promising solution could be the use of ensemble

methods combining oversampling, undersampling, and SMOTE, to strike a balance between the strengths of each technique. The need for the alternate model in the training and classification of transcript is because of the poor performance achieved in the prediction. The Bi-directional Gated Recurrent Unit has an F1 score of 0.36 for the depressed class. The performance in classification is low as the prime application of the overall model is to identify the depressed participants to be classified correctly from the trained model. The LSTM model worked slightly better compared to the earlier model. The performance might be poor due to the 5-flod cross validation as the stopping condition of EarlyStopping and learning rate reduction to avoid any overfitting during the training of the model.

In the audio modality, the training of the CNN model without any hyperparameter tuning classified the classes between depressed and non-depressed with an overall accuracy of 0.6. the hyperparameters used are 64 filters for each of the features fed into the convolutional layer, with a fixed kernel size of 3, dropout rate of 0.5 and learning rate of 0.001 compiled by the Adam optimizer. Compared to the model where hyper parameters tuned, with fixed filters of 128 and kernel size of 3 and the other parameters tuned, the model achieved an accuracy of 0.69 with the depressed class F1 score of 0.48.

In the initial stages of research, several techniques were tested but failed to yield satisfactory results. Classical machine learning algorithms, such as Gaussian Naive Bayes and logistic regression, were explored for audio feature classification but produced poor accuracy and precision. These models struggled due to the redundant nature of some of the selected features, weakening the overall performance.

Another approach involving direct spectrograms of audio files also failed to capture the frequency content effectively. This method was replaced by using Mel-spectrograms, which offered a more detailed representation of audio signals and contributed significantly to the improved accuracy of the CNN model.

Efforts were made to integrate the audio and text data using a gating mechanism, but this approach proved overly complex. Matching text and audio segments at precise time intervals introduced implementation challenges, leading to the adoption of separate deep learning models for each modality. The final multimodal fusion method, using a meta-learner, successfully overcame these challenges, achieving much better performance.

To concatenate the predictions from the two modalities, the better performing models from these 2 features are identified and trained. Using a Meta learner the model is trained using the prediction results from these two trained models. For this research, an attempt to use Logistic regression and Gradient boosting classifier is identified to make the total prediction using the Meta learner. In combining the prediction from GRU and tuned CNN model, the accuracy of the model is 0.63, with the F1 score of 0.38 for the depressed class. The similar model using the Gradient boosting classifier has an overall accuracy of 0.86 with an F1 score of 0.83 for the depressed class.

However, using the gradient boosting classifier for the Bi-LSTM and CNN, the overall accuracy for the classification of the depressed and non-depressed participants increases to 0.94 and an F1 score of 0.92 for the depressed class and the model has classified accurately the true positive and true negative cases. Hence this model is chosen as the final model for the classification of the depressive participants from the DAIC WOZ dataset.

# 7 Conclusion and Future Work

This study aimed to enhance depression classification by leveraging meta-learning approach to combine processed transcript and audio data using Bidirectional LSTM and CNN. There is a significant improvement in classification accuracy in integrating the audio and text modality compared to unimodal approaches. While the LSTM model captures the features from the transcripts and the CNN model captures the features from the audio file, the overall model is robust to classify the depressive nature of the person. The final model has predicted with an overall accuracy of 0.94. The confusion matrix shows that the model has identified 21 of the 23 non-depressed class and all 12 depressed class. Though the model prediction is good, there is a possibility that the model has learned the training data very well where the possibility to the extent that it might not generalize well to unseen data. The size of the test data is very small and there is a risk of the models performing well because of the smaller sample size. This can be checked by applying cross validation by comparing train and validation performance of the overall model.

Future improvements can explore more advanced data augmentation techniques for both text and audio. For textual data, models like BERT or GPT can be employed to generate diverse training samples, providing a more robust understanding of depressive language. These models capture contextual dependencies more effectively than traditional embeddings like GloVe.

For audio data, additional augmentation techniques such as pitch shifting, time stretching, and adding noise can be explored to increase the variability of the audio features. This will enhance the model's ability to generalize across different speakers and environments. In addition, the use of ensemble methods for handling class imbalance, such as stacking oversampling and undersampling techniques (e.g., "oversampling-undersampling-SMOTE") can provide a more balanced dataset for training. This would mitigate the risk of overfitting while preserving the essential characteristics of the minority class (depressed cases).

The potential real-time use case of this model is in assisting psychiatrists and psychologists, where the result from the model can be an additional layer of insight, reducing the chances of misdiagnosis and in remote mental health monitoring where it can help in areas where mental health services is limited. The future work might consider the possibility of bias in the data. Predominantly depression affects more women than men and future work might include the layer to capture the gender as a factor in classifying the depressed classes.

Also validating the model with other datasets to validate the performance to generalizability and integrate the models with cross language and cross-cultural contexts.

# References


Birdsey, N., 2020. Integrating CBT and CFT within a case formulation approach to reduce depression and anxiety in an older adult with a complex mental and physical health history: a single case study. The Cognitive Behaviour Therapist, 13, p.e41.

Dinkel, H., Wu, M. and Yu, K., 2019. Text-based depression detection on sparse data. arXiv preprint arXiv:1904.05154.

Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S. and Traum, D.R., 2014, May. The distress analysis interview corpus of human and computer interviews. In LREC (pp. 3123-3128).

Iyortsuun, N.K., Kim, S.H., Yang, H.J., Kim, S.W. and Jhon, M., 2024. Additive Cross-Modal Attention Network (ACMA) for Depression Detection based on Audio and Textual Features. IEEE Access.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T. and Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders, 114(1-3), pp.163-173.

Lin, L., Chen, X., Shen, Y. and Zhang, L., 2020. Towards automatic depression detection: A BiLSTM/1D CNN-based model. Applied Sciences, 10(23), p.8701.

Ma, X., Yang, H., Chen, Q., Huang, D. and Wang, Y., 2016, October. Depaudionet: An efficient deep model for audio-based depression classification. In Proceedings of the 6th international workshop on audio/visual emotion challenge (pp. 35-42).

Milintsevich, K., Sirts, K. and Dias, G., 2023. Towards automatic text-based estimation of depression through symptom prediction. Brain Informatics, 10(1), p.4.

Murphy, J., Sweeney, M.R. and McGrane, B., 2020. Physical activity and sports participation in Irish adolescents and associations with anxiety, depression and mental wellbeing. Findings from the physical activity and wellbeing (paws) study. Physical activity and health, 4(1), pp.107-119.

Nadeem, M., 2016. Identifying depression on Twitter. arXiv preprint arXiv:1607.07384.

Pandey, A. and Vishwakarma, D.K., 2023. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Applied Soft Computing*, p.111206.

Sardari, S., Nakisa, B., Rastgoo, M.N. and Eklund, P., 2022. Audio based depression detection using Convolutional Autoencoder. Expert Systems with Applications, 189, p.116076.

Shen, Y., Yang, H. and Lin, L., 2022, May. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6247-6251). IEEE.

Solieman, H. and Pustozerov, E.A., 2021, January. The detection of depression using multimodal models based on text and voice quality features. In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) (pp. 1843-1848). IEEE.

Tejaswini, V., Sathya Babu, K. and Sahoo, B., 2024. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), pp.1-20.

Yalamanchili, B., Kota, N.S., Abbaraju, M.S., Nadella, V.S.S. and Alluri, S.V., 2020, February. Real-time acoustic based depression detection using machine learning techniques. In 2020 International conference on emerging trends in information technology and engineering (ic-ETITE) (pp. 1-6). IEEE.