

UNIVERSITY OF BAHRAIN
COLLEGE OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING



Detecting and Analyzing of Depression Using Natural Language Processing (NLP).

A Thesis Submitted in Partial Fulfillment of the Requirements for the M.Sc. Degree
in Artificial Intelligence Systems

Submitted by

Hawra Ahmed Al Mohssin

20160966

Supervised by

Prof. Ebrahim Abdulla Mattar

Kingdom of Bahrain

Nov 2024

Publications

Hawra A. Al Mohsin, and Mattar, E. A. (2023). “*Detecting depression using natural language processing (NLP) techniques,*” 7th IET Smart Cities Symposium (SCS 2023). doi: 10.1049/icp.2024.1034.

Abstract

Sadness could be a common and severe health condition that impacts how you are, feeling think, and carry on. It can cause a range of emotional and physical issues, lessening a person's capacity to operate in a way of life. The seriousness of sadness can shift, extending from gentle and brief scenes to extreme and persevering conditions. It regularly emerges from a blend of hereditary, natural, natural, and mental variables. Certain hazard components, such as a family history of misery, past injury, unremitting therapeutic conditions, or substance mishandle, can increase defenselessness to the clutter. Early detection is key to successful treatment and superior results. Diagnosing misery more often than not includes a combination of self-reported side effects, clinical assessments, and perceptions of behavior, together with natural markers and advanced innovations. Innovative progressions have driven the improvement of advanced instruments and applications to help in identifying discouragement. These can incorporate versatile apps that utilize normal dialect handling (NLP) and machine learning to analyze content or discourse for depressive signs, wearable gadgets that track physiological temperament pointers, and online stages advertising self-assessment surveys. The proposal examines the utilize of machine learning models to distinguish discouragement through sound and content information from the eminent DAIC-WOZ database. This database incorporates clinical interviews outlined to facilitate the conclusion of mental trouble. The think about tackles highlights from the Collaborative Voice Investigation Store (COVAREP) and transformer innovations inside the dataset. It applies different classification models, such as CNN, LSTM, Arbitrary Forest, Support Vector Machine (SVM) Convolutional Recurrent Neural Network (CRNN) , and common dialect preparing (BERT) apparatuses. After broad testing and assessment, the CRNN show demonstrates to be the foremost successful in foreseeing sadness from sound information. Its prevalent execution highlights its unwavering quality as a instrument for discouragement location amid clinical evaluations. This investigation underlines the adequacy of CRNN in utilizing sound highlights for discouragement acknowledgment and propels understanding of how machine learning can be viably coordinated into mental health diagnostics.

Acknowledgments

"Whoever does not thank the creature does not thank the Creator"

Firstly, all praise and thanks to Allah Almighty for His abundant blessings and guidance that helped me complete this Thesis with ease, precision, and success. I extend my gratitude and appreciation to Professor Ebrahim Abdulla Mattar for trusting me and supporting me to accomplish this thesis to the fullest, never sparing any information or encouragement.

I also offer my sincere thanks to the University of Bahrain for giving me the opportunity to join the program and providing me with all the resources I needed to obtain a master's degree.

I dedicate this work to those who are the fuel of my life, those who believed in me and granted me encouragement, trust, support, and love; to my entire family (especially my late father, who left this world hoping I would reach the highest levels of knowledge and success) and to my mother who surrounds me with her prayers morning and evening; to all my brothers and sisters (Nawaf, Mufeed, Abdullah, Fatimah, Azhar, Mohammed, Mustafa) and my dear husband (Dr. Mohammed), and all my family members and friends for their support and encouragement during the writing of my thesis. Without their support, I would not have been able to reach the end.

TABLE OF CONTENTS

Defense committee’s approval.....	
Publications.....	02
Abstract.....	03
Dedication.....	
Acknowledgments.....	04
Table of Contents.....	05
List of Figures.....	08
List of Tables.....	09

CHAPTER ONE: INTRODUCTION TO THE STUDY

1.1. Motivation.....	13
1.2. Problem Statement.....	14
1.3. Objectives of The Study.....	15
1.4. Research Questions.....	15
1.5. Model of The Study.....	16
1.6. Limitations of The Study.....	16
1.7. Thesis Structure.....	17

CHAPTER TWO: STUDY BACKGROUND

2.1. Machine Learning.....	20
2.1.1. Supervised Learning.....	21
2.1.2. Unsupervised Learning.....	23
2.1.3. Semi-supervised Learning.....	25
2.1.4. Reinforcement Learning.....	26
2.2. Deep Learning.....	28
2.2.1. CNN.....	29
2.2.2. RNN.....	35
2.2.3. CRNN.....	35
2.1.4. Transformers.....	37
2.2. NLP.....	38
2.3. Mental Health.....	41
2.3.1. Depression.....	41

2.3.2. PHQ-8	44
2.4. Sound.....	46
2.4.1. Sound Attribute	46
2.4.2 Speech	50
2.4.3 Speech in Depression Detection	52
2.4.4 Audio Attributes.....	53

CHAPTER THREE: LITERATURE REVIEW

3.1. Introduction	59
3.2. Acoustic Features in Voice Analysis for Depression Detection	62
3.3. Machine Learning Approaches for Voice Analysis: Supervised Learning with SVMs.....	64
3.4. Examples of Voice-based Detection Systems	67
3.5. Depression Detection Models	69
3.5.1. Studies Exploring Depression Detection	69
3.5.2. Challenges and Considerations in Depression Detection	69
3.5.2.1. Data Privacy Concerns	69
3.5.2.2. Model Generalizability Issues	70
3.5.2.3. Role as Screening Tools	71
3.5.3. Future Directions in Machine Learning for Depression Detection.....	72
3.5.3.1. The Imperative of Explainable AI (XAI)	72
3.5.3.2. Multimodal Detection for an integral View	73
3.5.3.3. Addressing Diversity and Generalizability	74
3.6. Summary Table of Related work.....	76

CHAPTER FOUR: DEPRESSION DETECTION METHODOLOGY

4.1. Introduction	81
4.2. Dataset source.....	82
4.3. Feature Extraction	85
4.4. Classifiers	86
4.4.1. CNN	86
4.4.2. RF.....	87
4.4.3. SVM.....	88
4.4.4. LSTM.....	90

4.4.5.	BERT	92
4.5.	Evaluation Measures	96
4.6.	Data Preparation	98
4.7.	Audio Models	100
4.7.1.	CNN	100
4.7.2.	LSTM.....	102
4.7.2.1.	Model Configuration	103
4.7.2.2.	LSTM Comparison.....	103
4.7.2.3.	Architectural Difference	104
4.7.2.4.	Performance Results	105
4.7.2.5.	Analysis of Performance	108
4.7.3.	RF Classifier	109
4.7.4.	SVM Classifier.....	109
4.7.5.	CRNN Classifier	109
4.8.	Text Models.....	113
4.8.1.	Text Data Preprocessing	113
4.8.2.	BERT	114
4.8.3.	LSTM Applied to Text.....	116
4.8.3.1.	LSTM Model.....	116
4.8.3.2.	LSTM Comparison.....	119
4.8.3.3.	Architectural Difference	119
4.8.4.	CRNN Model in Text.....	121
4.8.4.1.	Performance and Training	122
4.8.3.2.	LSTM Comparison.....	119
4.8.3.3.	Architectural Difference	119

CHAPTER FIVE

RESULTS AND DISCUSSION

5.1.	Overview	126
5.1.1.	CNN Audio Models	126
5.1.2.	LSTM Models.....	126
5.1.3.	Random Forest and SVM Classifiers.....	127
5.1.4.	BERT Text Classification	127
5.1.5.	Comparison of LSTM Models on COVAREP Data.....	127

CHAPTER SIX

CONCLUSION AND FUTURE WORK:

6.1. Overview	131
6.2. Conclusion.....	131
6.3. Future work	132
References	141

LIST OF FIGURES

Figure 1 Typical neural network architecture	29
Figure 2 Example of cnn architecture for image classification.	30
Figure 3 The primary calculations executed at each step of convolutional layer.	32
Figure 4 Learning process for word2vec	39
Figure 5 Depression study region in KSA.	44
Figure 6 Distribution of PHQ-8 scores	46
Figure 7 Characteristics of sound wave	48
Figure 8 Decibel scale.....	49
Figure 9 Sound metrics measuring units.....	50
Figure 10 Representation of the vocal tract	51
Figure 11 Depression statistics	60
Figure 12 Representation of vocal tract	63
Figure 13 Representation of SVM.	65
Figure 14 Model workflow.	68
Figure 15 Selected models.	81
Figure 16 Study workflow.	81
Figure 17 Dataset content.	83
Figure 18 Data distribution of DAIC-WOZ depression database.....	84
Figure 19 Dataset splitting	85
Figure 20 CNN Architecture.....	87
Figure 21 Decoding Audio in the CNN Architecture.	87
Figure 22 Random Forest Algorithm.....	88
Figure 23 The Long Short-Term Memory (LSTM) Architecture.....	92
Figure 24 MLM structure of the BERT model.	94
Figure 25 BERT model.....	95
Figure 26 BERT training strategies	96

Figure 27 CNN model.....	101
Figure 28 CNN accuracy plot	101
Figure 29 CNN loss plot	101
Figure 30 CNN model classification report.....	102
Figure 31 LSTM model 1 summary.....	104
Figure 32 LSTM model 2 summary.....	105
Figure 33 LSTM model 1 accuracy plot	106
Figure 34 LSTM model 1 loss plot	106
Figure 35 LSTM model 1 classification report.....	106
Figure 36 LSTM model 2 accuracy plot	107
Figure 37 LSTM model 2 loss plot	107
Figure 38 LSTM model 2 classification report.....	108
Figure 39 RF model classification report.....	109
Figure 40 SVM model classification report.....	110
Figure 41 CRNN model Training	112
Figure 42 CRNN model Accuracy plot	112
Figure 43 CRNN model loss plot.....	112
Figure 44 CRNN model classification report	113
Figure 45 BERT model.....	115
Figure 46 BERT model accuracy plot	115
Figure 47 BERT model loss plot	116
Figure 48 LSTM text model	117
Figure 49 LSTM text model accuracy plot	118
Figure 50 LSTM text model loss plot	118
Figure 51 LSTM text model classification report.....	118
Figure 52 Basic LSTM text model.....	120
Figure 53 Basic LSTM text model classification report.....	120
Figure 54 CRNN model accuracy plot.....	123
Figure 55 CRNN model loss plot.....	124

List of Tables

Table 1 Comparative Study 1	62
Table 2 Comparative Study 2.	65
Table 3 Comparative Study 3.	68

Table 4 Summary of Related work. 76

Table 5 Train Development Test Set Attributes 84

Table 6 Confusion Matrix 98

Table 7 Models Performance Summary 126

List of Abbreviations

NLP	Natural Language Processing
CNN	Convolutional Neural Network
LSTM	Long short-term memory
SVM	Support Vector Machine
RF	Random Forest
KNN	Nearest Neighbors
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
PCA	Principal Component Analysis
ML	Machine Learning
DL	Deep Learning
t-SNE	t-Distributed Stochastic Neighbor Embedding
RL	Reinforcement learning
MDD	Major Depressive Disorder
SAD	Seasonal Affective Disorder
DSM-5-TR	Diagnostic and Statistical Manual for Mental Disorders
PHQ-8	Patient Health Questionnaire-8

CHAPTER ONE

STUDY INTRODUCTION

1.1 Motivation

Depressive clutter, commonly alluded to as discouragement, may be a predominant mental wellbeing condition characterized by tireless sentiments of pity or lack of engagement in exercises over amplified periods. Not at all like commonplace disposition variances, sadness can profoundly affect different aspects of life, counting connections, work, and scholastic execution.

Whereas misery can influence anybody, people who have experienced injury, critical misfortunes, or other upsetting occasions may be at a better hazard. Measurements demonstrate that ladies are more vulnerable to sadness than men, with an evaluated 3.8% of the worldwide populace influenced. Among grown-ups, around 5% encounter misery, with marginally higher rates watched among ladies than men. In addition, discouragement influences around 5.7% of grown-ups matured over 60.

On a worldwide scale, misery postures a noteworthy open wellbeing challenge, affecting over 280 million individuals around the world. Ladies are excessively influenced, with rates around 50% higher than those for men. Moreover, misery influences more than 10% of pregnant ladies and modern moms. Deplorably, suicide, regularly connected with sadness, claims the lives of over 700,000 people yearly, making it the fourth driving cause of passing among 15–29-year-olds. In spite of successful medications for mental disarranges being accessible, a significant parcel of people in moo- and middle-income nations need get to satisfactory care.

More than 75% of affected individuals in these locales don't get treatment due to different boundaries, counting lacking speculation in mental wellbeing care, a deficiency of prepared healthcare experts, and the disgrace encompassing mental sickness. Tending to these obstructions is pivotal for improving mental wellbeing administrations and lightening the worldwide burden of discouragement.

Distinguishing sadness involves recognizing its markers in people, which may show through discourse designs, dialect utilization, or facial expressions. Diagnosing sadness demonstrates challenging due to its differing introduction among people in terms of side effects, seriousness, and term.

For illustration, a few people may display disposition variances, whereas others may battle with decision-making or encounter increased uneasiness. Given the wealth of available online data, there exists noteworthy potential for conducting data-driven examinations and creating modern calculations to assist brain investigating masters in considering pity and refining clinical methodologies.

Determination of misery regularly involves considering different variables at the same time, such as side effects, therapeutic history, related clutters, and potential treatment choices, coming about in a time-consuming handle. Alternately, AI frameworks offer fast information handling capabilities, permitting for speedier determination suggestions based on treatment session information.

Precise location of sadness benefits from models that coordinated etymological data, vocal cues, and facial expressions, all of which contribute altogether to the acknowledgment prepare. Creating frameworks competent of understanding person behavioral designs and foreseeing misery status holds guarantee for the therapeutic community. This proposition centers on utilizing built up machine learning techniques and giving them sound and substance data to plan models for hopelessness acknowledgment. The machine learning models utilized include CNN, SVM, LSTM, RF, and NLP.

1.2 Problem Statement

In this investigation, the focus lies on executing machine/deep learning calculations and Normal Dialect Preparing methods to recognize discouragement from both human voice and content. This endeavor utilizes the eminent DAIC-WOZ Database, which comprises clinical interviews custom-made for diagnosing mental trouble conditions such as uneasiness, misery, and post-traumatic stretch clutter.

These interviews shape portion of a broader activity pointed at creating a computer specialist competent of conducting interviews with people and perceiving verbal and nonverbal prompts demonstrative of mental sickness (DeVault et al., 2014). The dataset envelops sound and video recordings alongside comprehensive survey reactions. Outstandingly, it incorporates information from Wizard-of-Oz interviews, where a

vivified virtual questioner named Ellie is controlled by a human questioner arranged in a partitioned room.

1.3 Objectives of The Study

The essential point of this inquiry is to investigate the achievability of utilizing manufactured insights strategies to decide whether a person is encountering discouragement. This includes creating numerous machine learning and Characteristic Dialect Preparing models utilizing the dataset given by DAIC-WOZ. DAIC-WOZ, curated by the College of California, was displayed in the midst of the Audio/Visual Feeling Challenge and Workshop in 2017 (AVEC 2017). This dataset comprises clinical interviews counting an examiner and 189 individuals, with the objective of looking at distinctive mental inconvenience conditions such as wretchedness, uneasiness, and post-traumatic extend clutter. Due to specialized imperatives, as it where information from these 189 members have been utilized in this ponder.

The taking after are the different destinations of this proposal work:

1. To investigate literature for related inquire about.
2. To form an AI-based analyzing framework with control that can be utilized to identify either the individual discouraged or non-depressed.
3. To create diverse models for recognizing sadness utilizing solid dataset source DAIC-WOZ.
4. With a center on voice and content information, there's potential for future integration with advanced advances found in portable phones or progressed gadgets like Alexa. This integration seems to help within the early determination of sadness.

1.4 Research Questions

Assessing the significance of employing artificial intelligence techniques for detecting depression:

- How effective are machine learning models in accurately detecting depression
- from voice recordings?

- Can machine learning architectures improve the performance of depression detection from voice and text data?
- How do different machine learning algorithms (SVM, CNN, LSTM, RF) and NLP techniques compare in their ability to detect depression from voice and text data?

1.5 Model of The Study

This study examines the challenge of detecting depression in people and addresses in detail the techniques used to determine whether a person has depression or not, as it is one of the psychological diseases that affects the behavior of individuals, the way they talk, social interaction with individuals, and general behaviors. In this research, Leveraging the established DAIC-WOZ database, which contains clinical interviews classified as audio, written, and video. Through this research, machine learning techniques and natural language programming techniques are applied and compared between them.

1.6 Limitations of The Study

Every research endeavor encounters various challenges that can influence the attainment of satisfactory outcomes. Some of the challenges confronting our study and similar research efforts include:

- 1- The challenge of interpreting ambiguous or nuanced language expressions. Depression can manifest differently across individuals, and language may not always provide clear indicators of the condition. As a result, machine learning models may struggle to accurately detect depression in cases where symptoms are subtle or context dependent.
- 2- Privacy concerns may arise when analyzing sensitive voice and text data for depression detection. Ensuring data security and confidentiality while collecting and processing such information is crucial to protect individuals' privacy rights and maintain trust in the system.

- 3- The performance of machine learning models for discouragement discovery may change depending on the quality and amount of the information accessible. Constrained or boisterous information can prevent the model's capacity to memorize important designs and lead to imperfect execution in real-world applications.

1.7 Thesis Structure

The proposition is organized into six chapters, each contributing to a comprehensive understanding of the ultimate results. Here's an overview of the chapters:

Chapter 2:

This chapter begins with an outline of machine learning methods and after that delves deeply into the topic of sadness, discussing its different forms. Statistical data is given to emphasize the centrality of this mental wellbeing condition. The chapter too covers the Persistent Wellbeing Survey, counting test questions and score calculation. Furthermore, it delves into the basics of sound, counting fundamental concepts of sound and discourse, and broadly analyzes sound properties and sound highlights significant to the acknowledgment prepare.

Chapter 3:

This chapter reviews previous studies on depression recognition, emphasizing the utilize of machine learning and normal dialect preparing connected to assorted information sources.

Chapter 4:

In this chapter, the dataset utilized for the thesis is introduced, complete with descriptive statistics to shed light on its structure. It goes on to offer a comprehensive examination of the depression recognition methods employed in the thesis. Moreover, it identifies the optimal parameters for various models including CNN, LSTM, Random Forest, SVM, and NLP techniques such as BERT and LSTM. The chapter concludes by presenting the results of the depression recognition efforts, accompanied by an in-depth discussion of the evaluation metrics for each algorithm employed.

Chapter 5:

This chapter presented the LSTM Comparison in Audio and Text Comparison.

Chapter 6:

The final chapter summarizes the discoveries of the thesis and gives suggestions for further research on depression recognition through sound analysis.

CHAPTER TWO

STUDY BACKGROUND

2.1 Machine Learning

The emergence of machine learning has transformed artificial intelligence (AI) and data-driven decision-making across various fields [1], [2]. The exponential growth of data and rapid technological advancements necessitate innovative methods to extract meaningful insights and patterns from large datasets [3]. Traditional rule-based programming often struggles to manage the complexity and variability of real-world data, making machine learning a paradigm shift that allows systems to learn, adapt, and autonomously improve over time [4], [5].

In specific domains, traditional decision-making processes are constrained by the complexity of data, dynamic environments, and the need for accurate and timely predictions [6], [7]. As a result, utilizing machine learning is crucial for enhancing decision-making, optimizing resources, and mitigating risks [8].

The roots of machine learning trace back to the 1950s and 1960s through the foundational work of Alan Turing, who theorized intelligent machines capable of learning, and Arthur Samuel, who introduced self-learning checker-playing programs [4], [9]. In the 1970s and 1980s, the field expanded with the growth of AI and expert systems, though computational limitations hindered progress [5]. Advances in decision trees, neural networks, and pattern recognition during this period laid the foundation for future developments [10], [11].

The 1990s saw renewed interest in neural networks driven by backpropagation algorithms and increased computational power, leading to progress in image and speech recognition [12]. Support vector machines (SVMs) and ensemble methods also gained popularity [13]. In the 2000s, the rise of the internet and data-driven applications spurred further machine learning research, focusing on probabilistic models, reinforcement learning, and Bayesian techniques [7], [14].

During the 2010s, deep learning became dominant, propelled by innovations in neural network architectures, training algorithms, and GPU technology [15], [2]. Deep learning achieved unparalleled performance in image classification, NLP, and gaming [8]. This decade also saw the democratization of machine learning through open-source tools, cloud platforms, and online education [1].

In the 2020s, machine learning continues to evolve with advancements in reinforcement learning, meta-learning, and explainable AI [16], [17]. Ethical issues such as bias, fairness, and transparency have become central topics, prompting efforts to develop interpretable AI systems [18], [19]. As machine learning integrates further into industries like healthcare, finance, and autonomous systems, its impact on technology and human interaction is set to grow [20], [21].

Machine learning algorithms are generally classified into four types: supervised, semi-supervised, unsupervised, and reinforcement learning [7], [14].

2.1.1 Supervised learning

Supervised learning is a critical paradigm in machine learning where algorithms learn from labeled data consisting of input-output pairs. The goal is to learn a mapping from inputs to outputs, enabling the algorithm to make predictions or decisions on new, unseen data [1], [3]. In supervised learning, the algorithm is trained on a dataset where each data point is associated with a label or target value. This approach includes two primary types: regression and classification [5].

The most common algorithms used in supervised learning include logistic regression, support vector machines (SVM), random forests, Naïve Bayes, decision trees, and neural networks [6], [7].

Regression

In regression tasks, the algorithm predicts continuous output values. Examples include predicting house prices based on features such as size, location, and number of rooms, or forecasting stock prices based on historical data [4]. Regression aims to uncover relationships between dependent and independent variables. The objective of regression modeling is to estimate a function that maps input variables to a continuous output variable [6]. After training, common evaluation metrics include Mean Squared Error (MSE), which measures the difference between predicted and actual values [7]:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (1)$$

Root Mean Squared Error (RMSE) extends MSE [7]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

Mean Absolute Error (MAE) computes the absolute difference between predicted and actual values. Similar to RMSE, the MAE score is also expressed in the same units as the output variable [7]

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (3)$$

Types of supervised regression algorithms include:

- 1- Linear Regression: A regression algorithm used to predict continuous output values based on input features. It fits a linear model to the data and learns coefficients that minimize the difference between predicted and actual output values [6]

$$y = B_0 + B_{1x1} + B_{2x2} + \dots + B_{n \times n} + \epsilon \quad (4)$$

- 2- Support Vector Regression (SVR): Support Vector Regression (SVR) applies the principles of Support Vector Machines (SVM) to regression tasks. It aims to find a hyperplane that fits the data within a specified margin of tolerance, minimizing the error outside this margin. SVR is effective for both linear and non-linear regression by using kernel functions, such as the radial basis function (RBF), to map data into higher dimensions. SVR is robust against outliers and is well-suited for small to medium-sized datasets with complex patterns [13].

Classification

Supervised learning classifiers categorize data based on labeled training datasets. These algorithms are essential for tasks such as spam detection, medical diagnosis, and handwritten digit recognition [1]. The classifier learns from past examples by mapping input features to output labels. Data is typically split into training and testing sets to evaluate model performance. The effectiveness of classifiers depends on the quality of labeled data and the algorithm's ability to generalize to unseen data [1], [5]. Some of the popular classifiers used in supervised learning:

Linear Classifiers – Logistic regression and linear SVM assume data can be separated by a straight line or hyperplane. Logistic regression estimates probabilities using a sigmoid function, which is ideal for binary classification tasks [8]. Linear SVMs maximize the margin between classes, reducing overfitting in high-dimensional spaces [13].

Non-Linear Classifiers – When data is not linearly separable, algorithms like kernel SVM, k-nearest neighbors (KNN), and decision trees are used. Kernel SVM maps data to higher-dimensional spaces, enabling non-linear decision boundaries [13]. KNN classifies new points based on the majority vote of neighboring data points, while decision trees recursively split data based on feature conditions [10].

Ensemble Methods – Techniques like random forests and gradient boosting improve accuracy by combining multiple models. Random forests aggregate predictions from several decision trees, reducing variance and improving generalization [6]. Gradient boosting refines weak learners iteratively, producing highly accurate models used in real-world applications [15].

Bayesian Classifiers – Naïve Bayes assumes feature independence and classifies data based on probability, making it efficient for text classification and spam detection. Bayesian networks model dependencies between features, handling uncertainty in data [7].

Neural Networks – Neural networks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), excel at handling large, unstructured datasets. CNNs are effective for image recognition, while RNNs process sequential data, such as speech and text [1], [2]. Deep learning techniques have significantly reduced the need for manual feature engineering [15].

2.1.2 Unsupervised learning

Unsupervised learning may well be a Portion of machine learning where the calculation learns plans and structures from unlabeled data, without unequivocal supervision or direction [1]. Not at all like managed learning, where the calculation is given with

labeled data (input-output sets), unsupervised learning deals with rough, unlabeled data and points to reveal covered up designs or groupings inside it [22].

1- Clustering: Clustering algorithms bunch comparative information focuses together based on a few closeness metrics. Common clustering calculations incorporate:

K-Means: Separates the information into k clusters by minimizing the within-cluster sum of squared distances [23].

Hierarchical Clustering: Builds a hierarchy of clusters by recursively blending or part clusters based on a closeness model [24].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Clusters dense regions of data points, separating areas of high density from areas of low density [25]

2- Dimensionality Reduction: Dimensionality reduction methods point to reduce the number of input highlights whereas protecting the foremost critical data. This makes a difference in visualizing high-dimensional information and lessening computational complexity [26]. Common dimensionality lessening calculations incorporate:

Principal Component Analysis (PCA): Linear dimensionality reduction procedure that finds the orthogonal tomahawks (central components) that maximize fluctuation within the information [27].

t-Distributed Stochastic Neighbor Embedding (t-SNE): Non-linear dimensionality decrease strategy that jam neighborhood likenesses between information focuses [28].

Autoencoders: Neural network-based procedure that learns to compress and decompress information, viably learning a lower-dimensional representation of the input information [29].

3- Anomaly Detection: Anomaly detection calculations distinguish information focuses that veer off altogether from the standard. These peculiarities might show

blunders, exceptions, or unordinary behavior [14]. Common peculiarity discovery strategies incorporate:

Density-Based Anomaly Detection: Recognizes irregularities as information focuses that lie in low-density locales of the information conveyance [25].

Isolation Forest: Constructs isolation trees to isolate anomalies efficiently by randomly partitioning the data space [30].

One-Class SVM: Learns a boundary around typical information data information. Synonyms focuses, classifying any focuses exterior the boundary as inconsistencies [31].

Unsupervised learning finds applications in different spaces, counting client division, exception discovery, picture and content clustering, and dimensionality diminishment [1]. It empowers investigation and disclosure in expansive datasets where names may be inaccessible or exorbitant to get. Unsupervised learning calculations play a significant part in information preprocessing, exploratory data analysis, and feature engineering, clearing the way for further insights and downstream analysis [22].

2.1.3 Semi-supervised learning

Semi-supervised learning may be a machine learning worldview that lies between supervised and unsupervised learning [32]. In supervised learning, the calculation is prepared on labeled information, where each input illustration is related with a comparing yield name. In unsupervised learning, the calculation is prepared on unlabeled data, where there are no yield names given [33]. Semi-supervised learning combines components of both approaches by utilizing a small sum of labeled information in conjunction with a bigger pool of unlabeled information for preparing [34].

The key thought behind semi-supervised learning is to leverage the abundant unlabeled information to progress the execution of the show prepared on restricted labeled information. This is often especially valuable in scenarios where getting labeled

information is costly or time-consuming, but unlabeled information is promptly accessible [35]. There are a few approaches to semi-supervised learning:

Self-training: The algorithm iteratively trains on the labeled information and after that employs the prepared show to anticipate names for the unlabeled information. The high-confidence forecasts are included to the labeled dataset, and the method is rehashed until joining [36].

Co-training: The algorithm trains multiple models on different subsets of highlights or seas of the information. Each model is at that point utilized to label the unlabeled data, and occurrences with consistent predictions from both models are added to the labeled dataset [37].

Semi-supervised Generative Models: These models learn a probabilistic representation of the data and utilize it to create names for the unlabeled examples [32].

Graph-based Methods: These methods represent the data as a chart, where hubs compare to examples and edges represent connections between them. Labels are propagated through the chart based on similitude measures or chart algorithms [38].

Semi-supervised learning has applications in different spaces, counting computer vision, characteristic dialect preparing, and discourse acknowledgment [33]. It permits for more proficient utilize of accessible information and can lead to made strides demonstrate execution, particularly when labeled information is rare [35]. Be that as it may, planning effective semi-supervised learning algorithms requires careful consideration of the fundamental data distribution and the relationships between labeled and unlabeled examples.

2.1.4 Reinforcement learning

Fortification learning (RL) may be a sort of machine learning worldview where a specialist learns to form choices by association with an environment [39]. Not at all like administered learning, where the calculation is ready on labeled data, or unsupervised learning, where the calculation learns plans from unlabeled data, fortification learning is based on the concept of trial and mistake [40].

In reinforcement learning, the specialist takes activities in an environment and gets input within the shape of rewards or disciplines based on the comes about of its exercises [41]. The objective of the master is to memorize a policy—a mapping from states to actions—that maximizes the aggregate remunerate over time [16]. The administrator learns through investigation, attempting out particular activities and watching their results, and misuse, leveraging its learned data to create choices that lead to higher rewards [15].

The key components of support learning are [39]:

Operator: The substance that interatomic with the environment and learns to create choices.

Environment: The outside framework with which the operator interatomic. It gives criticism to the specialist based on its activities.

State: A representation of the current circumstance or setting of the environment.

Activity: The choice made by the specialist to influence the state of the environment.

Remuneration: Input from the environment that demonstrates the attractive quality of the agent's activity. The RL prepare regularly includes the taking after steps [40]:

- **Initialization:** The specialist initializes its approach and other parameters.
- **Activity Determination:** The operator chooses an activity based on its current arrangement and the observed state of the environment Execution: The agent executes the chosen activity within the environment.
- **Perception:** The specialist gets criticism from the environment in the frame of a remunerate and watches the unused state.
- **Learning:** The operator overhauls its approach based on the watched state-action-reward moves, pointing to maximize aggregate rewards over time.

Support learning has applications in different spaces, counting mechanical autonomy, amusement playing, independent vehicles, proposal frameworks, and finance [18]. It is especially well-suited for problems where the operator must make consecutive choices in an energetic and questionable environment [42]. Be that as it may, RL

calculations can be challenging to prepare and require cautious tuning of parameters and compensate capacities to attain wanted behavior [17].

2.2 Deep Learning

Deep learning may be a specialized subset of machine learning that utilizes multi-layered neural systems to imitate complex designs and undertake an assortment of errands, counting picture acknowledgment, discourse acknowledgment, and common dialect handling. The foundational structure of deep learning is based on neural arrange models. Imitating the human brain, these systems are composed of interconnected units known as hubs or neurons, arranged in layers that interface inputs to yields. The layers situated between the input and yield are known as hidden layers. The descriptor "deep" in deep learning regularly alludes to the number of covered up layers displayed within the arrangement, with a few deep learning models highlighting hundreds or indeed thousands of these layers [1], [2], [11].

Deep learning models are created through the preparing of broad sets of labeled information, and they have the capability to specifically learn highlights from the data, eliminating the requirement for manual include extraction [6]. The concept of the primary artificial neural organize was presented in 1958, but the considerable computational control required for deep learning was not accessible until the 2000s [43], [15]. Today, with advanced computing resources, researchers can construct and train complex networks featuring extensive connectivity and numerous neurons [44].

The advent of high-performance GPUs has accelerated the development of deep learning models. The parallel architecture of GPUs is particularly well-suited for matrix operations and large-scale data processing, reducing training time from weeks to hours [45], [46]. When used in clusters or cloud computing environments, GPUs enable scalable and efficient deep learning, further driving advancements in fields like autonomous systems, medical diagnostics, and generative AI [47].

Deep learning encompasses various architectures, including:

1. **Convolutional Neural Networks (CNNs)** – Primarily used for image and video processing, CNNs apply convolutional filters to extract spatial features from input data [12]. CNNs are designed to recognize patterns through hierarchical

feature learning, making them highly effective in tasks such as facial recognition and medical imaging [48].

2. **Recurrent Neural Networks (RNNs)** – RNNs process sequential data by maintaining memory of previous inputs through hidden states. This architecture is widely applied in speech recognition and language modeling [49]. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are advanced forms of RNNs that address issues related to long-term dependencies and vanishing gradients [50].
3. **Convolutional Recurrent Neural Networks (CRNNs)** – CRNNs combine the feature extraction capabilities of CNNs with the sequential modeling power of RNNs. This hybrid model is used for tasks such as handwriting recognition, video analysis, and speech-to-text applications [51].
4. **Transformer Models** – Transformers, introduced by Vaswani et al. in 2017, are the foundation of modern NLP models. They rely on self-attention mechanisms to process entire sequences of data in parallel, significantly outperforming RNNs and CNNs in tasks like machine translation and text generation [52]. Transformers are the basis for models like BERT, GPT, and T5, which have set new benchmarks in AI research [53], [54].

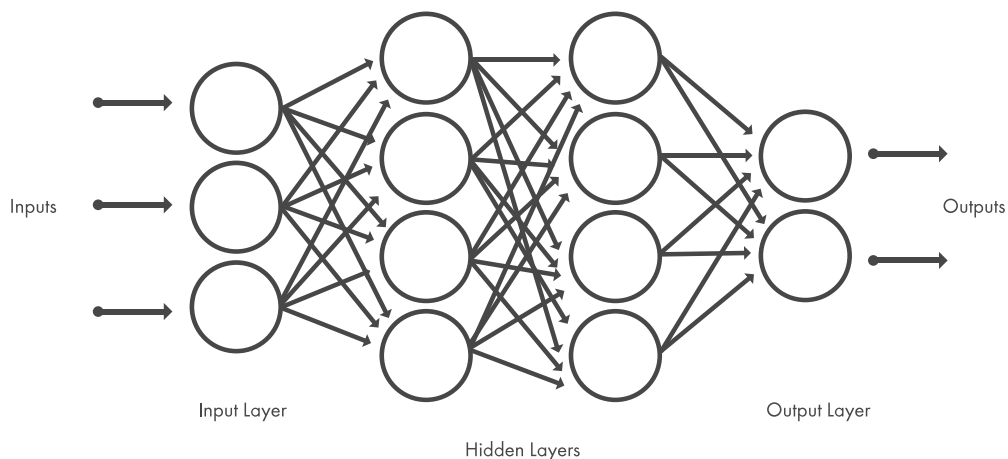


Figure 1: Typical neural network architecture [55]

2.2.1 CNN

Convolutional Neural Network (CNN) utilizes 2D convolutional layers to convolve learned highlights with input information, an approach that makes this architecture particularly effective for handling 2D data like pictures [15]. CNNs work by specifically extricating highlights from pictures, learning the significant highlights as the arrangement is prepared on a dataset of pictures. This handle of robotized highlight extraction improves the accuracy of deep learning models in picture classification assignments [12]. Past pictures, CNNs are also adaptable for classifying different types of information, such as time arrangement and text [1].

The primary advantage of Convolutional Neural Networks (CNN) over past models is their capacity to independently distinguish germane highlights without human intervention [40]. CNNs are widely utilized over different spaces such as computer vision, discourse preparing, and confront acknowledgment [56]. The plan of CNNs draws motivation from the neuronal structures within the brains of people and creatures, comparative to traditional neural systems [57]. For instance, the organization of cells in a cat's visual cortex, which processes visual information, is emulated by the architecture of a CNN [44].

CNNs offer several distinct benefits including equivalent representations, sparse interactions, and parameter sharing [58]. These networks utilize shared weights and local connections to effectively handle 2D data structures like picture signals, contrary to the fully connected networks that use many more parameters [59]. This approach decreases the number of parameters altogether, which disentangles the preparing handle and improves the network's speed, mirroring the function of cells within the visual cortex that as it were sense little parts of a scene, effectively capturing local correlations in the input as if applying local filters over it [60].

A typical CNN, akin to a Multi-Layer Perceptron (MLP), highlights numerous convolutional layers taken after by sub-sampling (pooling) layers, with completely associated layers concluding the design [61]. An example of a CNN setup for picture classification may well be depicted in a diagram such as Figure 2, illustrating this layered approach.

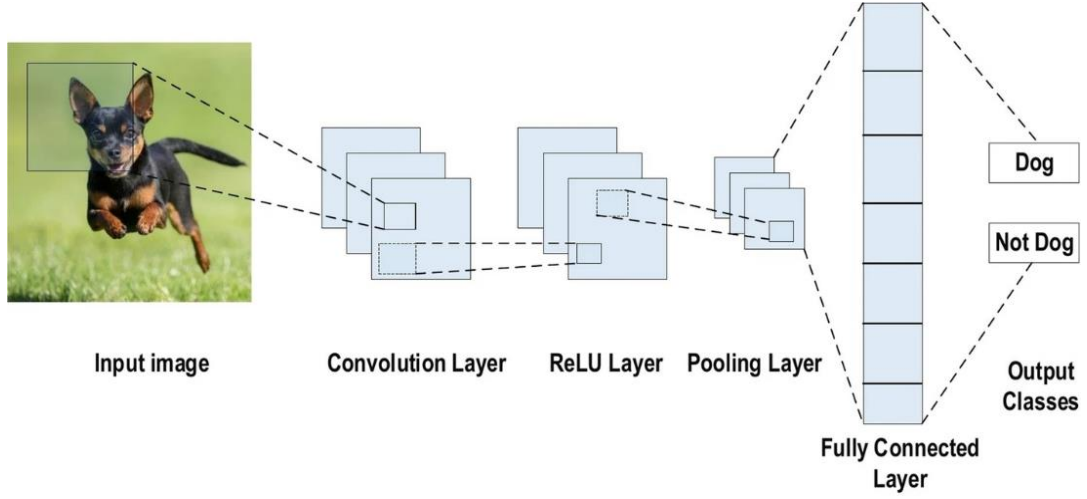


Figure 2: Example of CNN architecture for image classification [55]

In a CNN model, the input X for each layer is structured in three dimensions: height, width, and depth (channels), typically denoted as (m, m, r) . Here, height (m) is the same as the width, and the depth, often referred to as the channel number, represents different aspects of the input data, such as the color channels in an RGB image where $r=3$ [57].

Each convolutional layer contains multiple kernels (or filters), kk , which also have three dimensions $(n \times n \times q)$. These dimensions are constrained so that nn is smaller than mm , and qq is either equal to or smaller than r [62]. The kernels are central to establishing local connections within the network and share parameters—biases b^k and weights w^k —to generate k feature maps. Each feature map has dimensions $(m-n-1)$ after convolution [44].

The convolutional layer operates by computing a dot product between the kernel weights and local patches of the input—a process akin to that used in natural language processing (NLP) but applied to discrete segments of the input image. Following the convolution, an activation function, such as ReLU or sigmoid, is applied to introduce nonlinearity to the output of the convolution layer, resulting in the transformed feature maps. This sequence of operations forms the basic computation unit in a CNN, facilitating complex pattern recognition and feature extraction from the input data [63].

$$h^k = f(W^k * x + b^k) \quad (6)$$

The CNN models contain many layers, and each layer plays an important role in building the model.

1- **Convolutional Layer:** The convolutional layer is a fundamental element of CNN architecture. It is made up of a set of convolutional filters, often referred to as kernels. These kernels interact with the input image, which is represented as an N-dimensional matrix, through a process called convolution. This interaction produces an output known as a feature map, capturing important features from the input image [64].

- **Kernel:** A kernel in a CNN is essentially a grid of discrete values, each termed a kernel weight. Initially, these weights are assigned random values at the start of the CNN training process. Various methods exist for initializing these weights effectively. As training progresses through each epoch, these weights are adjusted, allowing the kernel to learn and extract important features from the input data [65].
- **Convolutional Operation:** The input to a CNN is typically a multi-channelled image, such as a grayscale image (single-channel) or an RGB image (three-channelled), unlike traditional neural networks that use vector format inputs [66]. To illustrate the convolutional operation, consider a grayscale image being processed by a kernel initialized with random weights. The kernel moves across the image both horizontally and vertically. At each position, a dot product is computed between the kernel and the overlaying area of the image; the values are multiplied correspondingly and then summed to produce a single scalar value. This process is repeated across the entire image until the kernel can no longer slide further. The series of scalar values obtained from these computations forms the feature map of the output. This is depicted in Figure 3, where the kernel is shown in light green, the corresponding area of the input image in light blue, and the resulting product values summed up are marked in light orange, indicating the contributions to the output feature map [67].

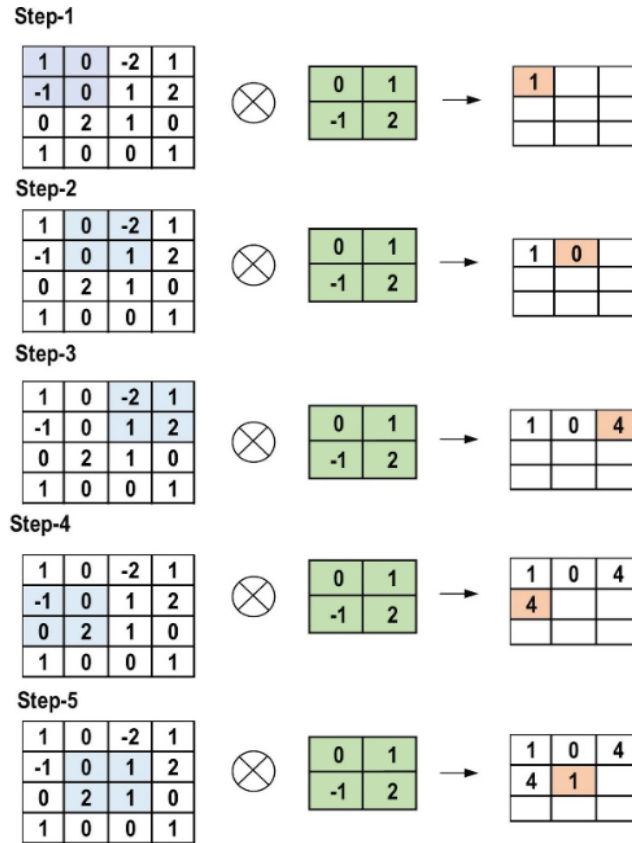


Figure 3: The primary calculations executed at each step of convolutional layer [55]

- 2- **Pooling Layer:** The primary function of the pooling layer is to down sample the feature maps produced by the convolutional operations. Essentially, this process reduces the dimensions of large feature maps to smaller ones while preserving most of the significant information or features [68]. Similar to the convolutional process, the size of the kernel and the stride are predefined before executing the pooling operation. There are several pooling techniques that can be applied, including max pooling, average pooling, min pooling, tree pooling, gated pooling, global average pooling (GAP), and global max pooling. The most commonly used methods among these are max pooling, min pooling, and GAP. These techniques help to reduce computational complexity and prevent overfitting by abstracting the higher-level features [69].
- 3- **Activation Function (Non-linearity):** The main role of the activation function in neural networks is to map the input to an output. This mapping is determined by computing a weighted summation of the inputs to a neuron, plus a bias term if it is included. The activation function thus decides whether a neuron should activate in response to the given input, producing an output accordingly.

In the architecture of a CNN, non-linear activation functions are placed after every layer that contains weights—commonly known as learnable layers. These include fully connected (FC) layers and convolutional layers. The introduction of non-linearity by these activation functions is crucial because it allows the network to map inputs to outputs in a non-linear manner, enabling CNN to learn more complex patterns and dependencies. Additionally, it is essential for activation functions to be differentiable. This property is vital because it allows the use of backpropagation—a method for effectively training the network by calculating gradients and updating the weights [70].

Commonly used activation functions in CNNs and other deep neural networks include:

- **ReLU (Rectified Linear Unit):** Activates a neuron only if the input is above zero, making it very efficient and widely used

$$f(x)_{ReLU} = \max(0, x) \quad (7)$$

- **Sigmoid:** Maps the input into a probability between 0 and 1, useful for binary classification

$$f(x)_{sigm} = \frac{1}{1+e^{-x}} \quad (8)$$

- **Tanh (Hyperbolic Tangent):** Similar to the sigmoid but maps values between -1 and 1

$$f(x)_{tanh} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (9)$$

- **Leaky ReLU:** A variant of ReLU that allows a small, non-zero gradient when the unit is inactive and can help maintain the flow of gradients throughout the network

These activation functions play a pivotal role in deep learning by adding the necessary complexity and depth to the model's learning capability

$$f(x)_{LeakyReLU} = \begin{cases} x, & \text{if } x > 0 \\ mx, & \text{if } x \leq 0 \end{cases} \quad (10)$$

4- **Fully Connected Layer:** Typically positioned towards the end of a CNN architecture, the fully connected (FC) layer features neurons that are interconnected

with every neuron from the preceding layer, embodying the fully connected approach. This layer acts as the classifier within CNN. It operates similarly to a traditional multilayer perceptron neural network, which is a type of feed-forward artificial neural network (ANN). The input to the FC layer is derived from the output of the final pooling or convolutional layer. This input is vectorized, a process achieved by flattening the feature maps into a single-dimensional array that feeds into the fully connected layer [71].

5- Loss functions: It plays a critical role in training Convolutional Neural Networks (CNNs) by evaluating the error between the predicted outputs (predictions) and the actual outputs (labels or targets). The primary goal of a loss function is to guide the training process by minimizing the error during the optimization of the network's weights [72]. Some common types of loss functions used in different scenarios:

- **Cross-Entropy or Softmax Loss Function:** This loss function is frequently used to assess the performance of CNN models in classification tasks and is also known as log loss. It outputs probabilities and is often used as an alternative to the mean squared error loss function for multi-class classification problems. In the output layer, it utilizes softmax activations to convert logits into a probability distribution across various classes [73]. The mathematical expression for the probability of each output class is typically presented as

$$p_i = \frac{e^{a_i}}{\sum_{k=1}^N e^{a_k}} \quad (11)$$

represents the unnormalized output from the preceding layer, with N denoting the number of neurons in the output layer. The mathematical expression for the cross-entropy loss function is given as

$$H(p, y) = - \sum_i y_i \log(p_i) \text{ where } i \in [1, N] \quad (12)$$

where y_i represents the true label, and p_i is the predicted probability for class i [52].

2.2.2 RNN

Recurrent Neural Networks (RNNs) are a type of neural network particularly that incorporates internal loops suited for processing sequences of data, making them ideal for tasks involving time series data, natural language processing, and more [49], [50].

These loops enable recursive dynamics within the network, creating delayed activation dependencies among the processing elements (PEs) [49]. Unlike most neural networks that use distributed representations—where information is encoded across the activations of multiple PEs—recurrent networks can also represent information through the time-varying activations of one or more PEs [15], [50]. As a result, RNNs can encode information both spatially across different PEs and temporally over time, which is why they are sometimes referred to as Spatio-Temporal Networks [49].

In an RNN, the flow of time is continuous, making the PE activations part of a dynamic system [15], [50]. This system can be described by a series of integral equations where the activations are determined by integrating past activations through a temporal kernel [49]. If we denote the activation of P_{ij} at time t as $y_i(t)$, and the temporal kernel as k , with f being a transformation function (commonly a sigmoid-type nonlinearity), then the activation dynamics can be expressed mathematically in this framework as:

$$y_i(t) = f \left(\sum_j \int_{t=0}^t k_{ij}(t' - t) \cdot y_j(t') \delta t' \right) \quad (13)$$

This formulation highlights how RNNs can model temporal dependencies, making them highly effective for sequence modeling tasks, such as machine translation, speech recognition, and video analysis [49], [74].

One of the most significant advantages of RNNs is their ability to handle variable-length input and output sequences. This makes them particularly useful in machine translation, where the length of input sentences may differ from the translated output. RNNs' flexibility allows them to adapt to diverse NLP tasks, from question answering to speech recognition [49]. Furthermore, their contextual learning approach enables RNNs to generate predictions based on the entire sequence of preceding words, improving accuracy in complex tasks like text summarization and document classification [52]. The ability to learn directly from raw text also reduces the need for extensive feature engineering, simplifying the development pipeline and accelerating model deployment [75].

Despite these strengths, RNNs face notable challenges, the most critical being the vanishing and exploding gradient problems. As gradients propagate backward during training, they can diminish or grow uncontrollably, making it difficult for the model to learn long-term dependencies. This results in poor performance when processing long sequences, as the network struggles to retain information from distant time steps [1]. In practice, this limits the effectiveness of RNNs in tasks requiring long-context understanding, such as document-level machine translation or long-form text generation [76].

Another limitation of RNNs is their bias toward recent inputs. The sequential nature of the architecture causes the network to prioritize information from the most recent time steps, often overlooking important contextual information from earlier parts of the sequence. This recency bias can negatively affect tasks like document classification, where context from the beginning of the text is crucial for accurate predictions [50]. Moreover, the sequential nature of RNN training prevents efficient parallelization, making them slower to train compared to models like Transformers, which process entire sequences simultaneously [74]. This limits their scalability for large datasets and complex NLP tasks.

To address these limitations, more advanced architectures such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) have been developed. These models introduce gating mechanisms that regulate the flow of information, mitigating the vanishing gradient problem and improving the model's ability to capture long-term dependencies [77]. Bidirectional RNNs (Bi-RNNs) enhance this further by processing data in both forward and backward directions, capturing context from both ends of the input sequence. Additionally, attention mechanisms, which selectively focus on relevant parts of the input, have significantly improved RNN performance, paving the way for the development of Transformer models that outperform RNNs in many NLP tasks [78].

2.2.3 CRNN

Convolutional Recurrent Neural Networks (CRNNs) are a hybrid deep learning architecture that combines Convolutional Neural Network (CNNs) and Recurrent Neural Networks (RNNs). The basic structure of CRNNs involves using CNN layers to

extract features from the data combined by RNN layers to capture the temporal relationships. This combination allows CRNNs to perform better than traditional models in many fields, including speech recognition, handwriting recognition, and video analysis. By combining the features of both Neural Network classes, CNNs and RNNs, CRNNs can effectively handle complex input data that exhibits both spatial hierarchies and temporal dependencies [79][80].

CNNs and RNNs both have made significant advances in deep learning. As mentioned earlier, CNNs excels at recognizing spatial patterns in data making them highly effective for tasks like image processing. On the other hand, RNNs are designed to handle sequential data making them ideal for tasks like language processing and time series prediction. However, when the task requires both spatial and temporal information, each class alone might be insufficient [80].

CRNNs address these limitations by integrating the strengths of both classes into a powerful and versatile class of neural networks. This integration enables CRNNs to perform well in tasks like recognizing cursive handwriting in natural scenes or predicting traffic patterns, where understanding both the structure and sequence of the data is important [81].

CRNN operate by first using CNN layers to extract spatial features from the input data. For instance, in the case of image data, the CNN layers identify important patterns such as edges, shapes, and textures, reducing the data's dimensionality while obtaining critical information. These extracted features are then passed to the RNN layers, which are designed to capture the temporal dependencies or sequences within the data. The RNN layers, typically implemented as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks, process the sequential information to maintain an understanding of how the data changes over time. The final output of the RNN layers is usually fed into a fully connected (dense) layer, followed by a softmax activation function, to produce a probability distribution over the possible classes. The class with the highest probability is selected as the output of the classifier [82].

One of the key benefits of CRNNs is their ability to learn from raw data directly, without needing extensive manual feature extraction. This makes CRNNs highly adaptable to various types of sequential data and different applications. Their ability to

perform end-to-end learning allows CRNNs to achieve top performance in tasks that require both spatial and temporal analysis [80].

Despite their strengths, CRNNs come with certain challenges, mainly related to the complexity of training and the resources required. The combined CNN and RNN layers mean CRNNs have a large number of parameters, which can make training computationally expensive and time-consuming. Additionally, CRNNs can suffer from issues like vanishing gradients, particularly in the RNN layers, which can hinder the learning process [79].

Another challenge is the difficulty in interpreting the outputs of CRNN models. The hybrid nature of CRNNs makes it hard to understand how the CNN and RNN parts contribute to the final decision, which can be problematic in applications like medical diagnosis. To overcome these challenges, future research may focus on optimizing CRNN architectures to reduce their computational demands and improve their interpretability. Techniques like model pruning or integrating attention mechanisms could help make CRNNs more efficient and easier to understand. As these improvements are made, CRNNs are expected to become even more widely used in areas such as real-time speech translation and video analysis [79][83].

2.2.4 Transformers

Transformers represent a fundamental shift in neural network architecture, addressing the limitations of traditional sequential models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. Unlike these models, which process data sequentially, transformers utilize a self-attention mechanism that allows them to analyze entire input sequences in parallel. This design enables transformers to capture dependencies between distant elements in the data, making them particularly effective for tasks involving long-range dependencies. This capability is crucial for natural language processing (NLP), where the meaning of a word often depends on the context provided by distant words in a sentence. As a result, transformers have become the foundation for large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer),

which have set new benchmarks in machine translation, text generation, and question answering [52][53].

The key innovation behind transformers lies in the self-attention mechanism, which dynamically weighs the importance of different words in a sequence. For a given input sequence $X = [x_1, x_2, \dots, x_n]$, the model first projects the input embeddings into three matrices: query (Q), key (K), and value (V) representations. These projections are computed as:

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v$$

where W_q, W_k, W_v are learnable weight matrices. The attention scores are obtained by computing the scaled dot product between Q and K , normalized by the square root of the dimension d_k :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

This normalization prevents the gradients from becoming excessively large, ensuring stable training. The softmax function converts the attention scores into a probability distribution, highlighting the most relevant tokens for each position in the sequence [47].

To enhance the model's capacity to learn different relationships, transformers employ multi-head attention, where multiple attention mechanisms operate in parallel. Each attention head attends to different parts of the sequence, and their outputs are concatenated and linearly projected back into the input space:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot W_o$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(QW_{q_i}, KW_{k_i}, VW_{v_i})$$

This mechanism allows the model to learn diverse representations at different levels of abstraction, making it robust across various NLP tasks [78].

A critical aspect of the transformer architecture is its ability to retain the positional information of tokens, as self-attention alone does not inherently encode sequence order. To address this, positional encodings are added to the input embeddings:

$$P_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad P_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

where pos is the token position and i denotes the embedding dimension. These sinusoidal encodings allow the model to differentiate tokens based on their absolute and relative positions, preserving sequential information even when inputs are processed in parallel [52].

The encoder portion of the transformer, as exemplified by BERT, consists of multiple identical layers, each comprising a multi-head self-attention mechanism followed by a feedforward neural network (FFN). The FFN is applied independently to each token:

$$\text{FFN}(x) = \max(0, xW_1 + b_1) \cdot W_2 + b_2$$

This two-layer feedforward network introduces non-linearity, enabling the model to learn complex mappings. Layer normalization and residual connections further stabilize training and improve gradient flow:

$$Z = \text{LayerNorm}(x + \text{Attention}(x))$$

$$\text{Output} = \text{LayerNorm}(Z + \text{FFN}(Z))$$

These components collectively enhance the model’s ability to generalize across tasks, leading to state-of-the-art results in various NLP benchmarks [71].

In contrast, the decoder, as used in models like GPT, follows a similar structure but incorporates masked self-attention to ensure that predictions for a given token only depend on previous tokens, preserving autoregressive behavior. This masking prevents information leakage from future tokens by setting the attention scores of future positions to $-\infty$:

$$M_{ij} = \begin{cases} 0, & \text{if } j \leq i \\ -\infty, & \text{if } j > i \end{cases}$$

The decoder also includes cross-attention layers, which attend to the output of the encoder, allowing the model to align and generate text based on input sequences. This encoder-decoder structure forms the backbone of machine translation systems such as T5 (Text-to-Text Transfer Transformer) and MarianMT [84].

The widespread adoption of transformers has catalyzed progress not only in NLP but also in computer vision and bioinformatics. Vision Transformers (ViTs) apply self-

attention to image patches, achieving performance comparable to convolutional neural networks (CNNs) on image classification tasks. Similarly, AlphaFold, a breakthrough in protein structure prediction, leverages transformers to model protein sequences and interactions [85].

Despite their successes, transformers present challenges, including substantial computational costs and data requirements. Training large models such as GPT-4 demands immense hardware resources, raising concerns about accessibility and sustainability. Techniques such as sparse transformers and distillation have been proposed to reduce the computational burden, enabling broader deployment without compromising performance [86].

2.3 NLP

Natural Language Processing (NLP) is a field within artificial intelligence (AI) that focuses on the interaction between computers and human languages. It involves developing algorithms and models that allow machines to understand, interpret, and generate human language in a way that is both meaningful and useful. It bridges the gap between human communication and machine understanding, enabling applications like language translation, sentiment analysis, speech recognition, and chatbots [87].

NLP uses a combination of linguistics, computer science, and machine learning techniques to process and analyze large amounts of natural language data. It involves tasks such as tokenization (breaking down text into smaller units like words or sentences), part-of-speech tagging (identifying the grammatical roles of words), and parsing (analyzing the grammatical structure of sentences). These foundational tasks provide users with more complex applications, such as text classification, information extraction, and machine translation [88].

The history of NLP goes back to the early days of computing, when simple rule-based systems were used to process text. These systems relied heavily on predefined grammar rules and dictionaries, limiting their capabilities. As computing power developed, and more data became available, statistical methods began to dominate the field. These methods, which involved using probabilistic models to predict language patterns, marked a significant shift towards more flexible and scalable NLP systems [87][89].

A major development in NLP came with the advancement of machine learning, especially deep learning, which allowed for the automatic learning of language representations from large datasets. Techniques like word embeddings became fundamental in modern NLP. The popular Word2Vec model, for example, learns word vectors by predicting the context in which words appear, based on the following objective function:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m; j \neq 0} \log P(w_{t+j} | w_t) \quad (14)$$

Where T is the total number of words in the corpus, m is the size of the context window, w_t is the target word, and w_{t+j} are the context words.

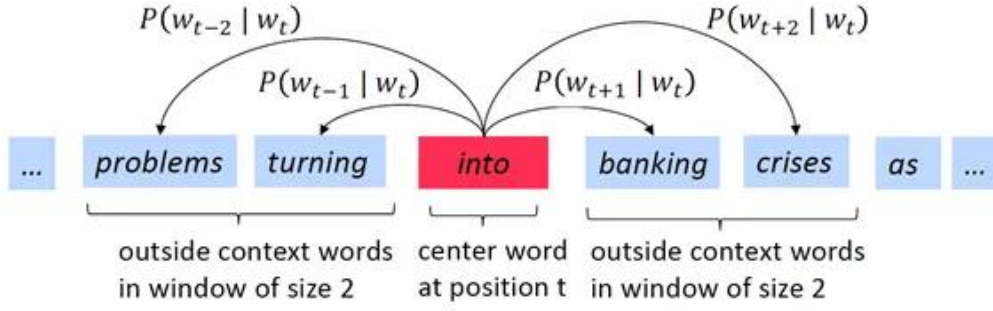


Figure 4: Learning process for Word2Vec [90]

The probability $P(w_{t+j}|w_t)$ is modeled using the softmax function:

$$P(w_{t+j}|w_t) = \frac{\exp(v'_{w_{t+j}} \cdot v_{w_t})}{\sum_{w \in V} \exp(v'_w \cdot v_{w_t})} \quad (15)$$

where v_{w_t} and $v'_{w_{t+j}}$ are the word and context word vectors, respectively, and V is the vocabulary [91].

Modern NLP systems rely heavily on deep learning models, particularly those based on neural networks. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer models, have become the standard models for processing sequential data in NLP tasks [1].

One of the critical component in modern NLP is the use of pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). These models are trained on great amounts of text data and fine-tuned for specific tasks, making them incredibly powerful and versatile. BERT, for instance, uses a bidirectional approach to understand the context of a word based on both its preceding and following words [53].

Despite the advancements in NLP, there are still significant challenges. One of the fundamental challenges is the handling of ambiguity and complexity in language. Words often have multiple meanings depending on the context, and understanding the exact meaning is crucial for accurate language processing [92]. Another challenge is the development of models that can generalize across different languages and dialects, as most NLP models are heavily biased towards English and other widely spoken languages [93].

Additionally, ethical concerns such as bias in language models and the generation of harmful content need to be addressed. As NLP models become more powerful, there is an increasing need for transparency and accountability in their development and deployment [94].

NLP has a wide range of applications across various industries and domains:

Information Retrieval: NLP techniques are used in search engines to understand user queries and retrieve relevant information from large text corpora.

Sentiment Analysis: NLP is employed to analyze the sentiment or opinion expressed in text data, such as social media posts or customer reviews, to gauge public opinion or customer satisfaction.

Machine Translation: NLP powers machine translation systems that enable automatic translation of text between different languages, such as Google Translate.

Speech Recognition: NLP algorithms are used in speech recognition systems to transcribe spoken language into text, enabling voice commands in virtual assistants like Siri or Alexa.

Text Summarization: NLP techniques are applied to automatically summarize large volumes of text, making it easier for users to extract key information from documents or articles.

Named Entity Recognition: NLP is used to identify and classify named entities such as people, organizations, or locations mentioned in text data, which is useful for tasks like information extraction or document categorization.

Question Answering: NLP powers question-answering systems that can understand user questions and provide relevant answers by extracting information from text data.

Dialogue Systems: NLP enables the development of conversational agents or chatbots that can engage in natural language conversations with users, providing assistance or answering queries.

Overall, NLP plays a crucial role in enabling computers to understand and interact with human language, opening up a wide range of possibilities for applications in communication, information processing, and automation. As NLP techniques continue to advance, we can expect to see even more sophisticated language-based AI systems in the future.

2.4 Mental health

2.4.1 Depression

Depression, or Major Depressive Disorder (MDD), is a significant mental health condition that adversely affects an individual's daily life. It is both common and severe, influencing emotional, physical, and psychological well-being [95]. Those affected by depression often experience persistent sadness, anxiety, diminished interest in activities they previously enjoyed, and low self-esteem [96]. Common symptoms include sleep disturbances, pessimistic views about the future, and even suicidal thoughts [97]. Additionally, individuals may exhibit slowed thinking, speech, or movement. Although these symptoms may resemble those of sadness or grief, depression is distinct from the normal grieving process, such as that experienced after the loss of a loved one or a job [98].

Depression is broadly categorized into two types in the Diagnostic and Statistical Manual for Mental Disorders (DSM-5-TR): Bipolar Disorder and Unipolar Depression [99]. Despite its name, Bipolar Disorder is included because it can involve depressing episodes. It is characterized by extreme mood swings ranging from high (manic episodes) to low (depressive phases), and sometimes these occur simultaneously in what are known as mixed episodes [100]. Bipolar Disorder is further divided into three categories:

- 1- **Bipolar I Disorder:** Marked by at least one manic episode that lasts a week or necessitates hospitalization [101].
- 2- **Bipolar II Disorder:** Involves a pattern of depressive episodes and hypomanic periods [101].

3- Cyclothymic Disorder or Cyclothymia: Features periods of hypomanic symptoms and depressive episodes lasting at least two years [99].

Unipolar Depression, primarily concerned with depressive "lows," requires at least five of the following symptoms during a two-week period, with one being either a depressed mood or loss of interest/pleasure [102]:

- 1- Depressed mood or irritability (in children and adolescents).
- 2- Diminished interest or pleasure in activities.
- 3- Significant weight change or appetite alteration.
- 4- Sleep disturbances.
- 5- Noticeable psychomotor changes.
- 6- Fatigue or loss of energy.
- 7- Feelings of worthlessness or excessive guilt.
- 8- Impaired ability to think or concentrate.
- 9- Recurrent thoughts of death or suicidal ideation.

Subtypes of MDD, known as specifiers, are defined based on duration and specific characteristics. Notable examples include:

1. **Persistent Depressive Disorder (Dysthymia):** A chronic form of depression lasting at least two years, often less intense than MDD but can include symptoms like low self-esteem and hopelessness [103].
2. **Seasonal Affective Disorder (SAD):** Depression that occurs seasonally, typically during less sunny months, must be consistent for at least two years to be diagnosed, and is characterized by periods of significant mood changes [104].

Postpartum Depression

In spite of the fact that ladies are at higher chance for common sadness, they are moreover at chance for postpartum misery, too called perinatal discouragement, which is impacted by regenerative hormones [105]. Ladies amid pregnancy or within the to begin with 12 months after conveyance endure from major or minor misery scenes [106]. Indications incorporate feeling discouraged most of the days, feeling far off from

family and companions, misfortune of interface, feeling tired or bothered and having sentiments of uneasiness or freeze assaults [96].

Psychotic Depression

This sort of discouragement can be created when a individual has extreme discouragement scenes in expansion to a few frame of psychosis, like daydreams or visualizations [107]. Discouragement can influence anybody, at any age, but it regularly starts at adulthood [108]. Hence, understanding and assessing the sadness predominance and recognizing the related hazard variables are fundamental for early location, make strides care, and direct future investigate in this populace [109]. In expansion, orderly prove collection is vital and can refine such prove in arrange to encourage depressive clutter avoidance and advancement programs among this cohort [110]. The figure appears the predominance of discouragement in cities within the Kingdom of Saudi Arabia in 2023 [111].

Depression, while undoubtedly difficult to endure, is highly treatable, with approximately 80% to 90% of individuals eventually responding positively to treatment [112]. Even severe cases can be effectively managed with appropriate interventions [113]. The treatment approach varies depending on the severity of the disorder, with early intervention being crucial for maximizing effectiveness [114].

Treatment regularly includes psychotherapy, pharmaceutical, or a combination of both [115]. Psychotherapy, regularly known as "conversation treatment," joins procedures such as cognitive behavioral treatment (CBT) and interpersonal treatment (IPT) [116]. Upper medicines may moreover be endorsed to direct brain chemicals capable for temperament and push control [117]. In cases where other medications demonstrate ineffectuality, brain incitement treatments like electroconvulsive treatment (ECT) may be considered [118].

In addition to professional interventions, lifestyle adjustments and changes in mindset can complement the treatment process [119]. Simple lifestyle modifications such as regular exercise and setting realistic goals can significantly contribute to managing depression symptoms and improving overall well-being [120]. By adopting a comprehensive approach that addresses both psychological and lifestyle factors, individuals can enhance their resilience and facilitate their journey towards recovery from depression [121].

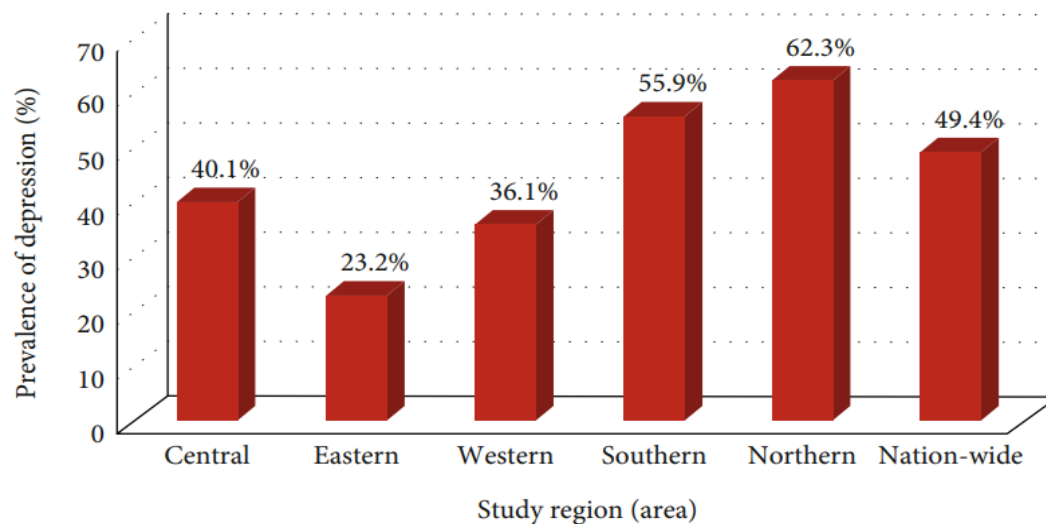


Figure 5: Depression study region in KSA [49]

2.3.3 PHQ-8

The Patient Health Questionnaire-8 (PHQ-8) is a self-administered diagnostic tool used to evaluate the severity of depression symptoms in individuals aged 12 and older [123]. It consists of eight items that reflect the criteria for diagnosing Major Depressive Disorder (MDD) as outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [96]. The PHQ-8 is similar to the PHQ-9 but omits the item on suicidal ideation for ethical reasons [124].

The respondent of PHQ-8 answer each question based on how many days in the past two weeks they have experienced specific symptoms of depression [125]. The responses are categorized as: 0-1 days="not at all", 2-6 days="several days", 7-11

days="more than half the days", and 12-14 days="nearly every day", each assigned a score from 0 to 3. The questions include:

- 1- Little interested or bliss in doing things.
- 2- Feeling down, disheartened, or pitiful
- 3- Burden falling or remaining resting, or resting as well much
- 4- Feeling tired or having little imperativeness
- 5- Dejected longing for or gorging
- 6- Feeling terrible nearly yourself – or essentially are a disappointment or have let yourself or your family down.
- 7- Bother concentrating on things, such as scrutinizing the everyday paper or watching tv
- 8- Moving or talking so gradually that other people may not have taken note. Or the converse – being anxious or energetic simply fair have been moving around a portion more than customary.

The total score, which ranges from 0 to 24, is obtained by adding up the scores of each question as shown in Figure 6 indicating the severity of depressive symptoms: 0 to 4 suggests no significant symptoms, 5 to 9 mild, 10 to 14 moderate, 15 to 19 moderately severe, and 20 to 24 severe. Diagnosis can be confirmed if either the first or second item is scored as "more than half the days" along with at least five of the other symptoms falling into the same frequency category, or by achieving a PHQ-8 score of 10 or higher [126].

The PHQ-8 is widely used in clinical settings as an effective and brief tool for screening and monitoring depression [127]. It assists healthcare providers in identifying individuals who might need more detailed evaluation or treatment for depression [124]. Additionally, it can be employed to monitor changes in symptoms over time and evaluate the effectiveness of therapeutic interventions [128].

Over the **last 2 weeks**, how often have you been bothered by any of the following problems?
(circle **one** number on each line)

How often during the past 2 weeks were you bothered by...	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

Figure 6: Distribution of PHQ-8 scores [123]

2.4 Sound

Sound plays a crucial role in human communication, allowing us to communicate through speech, music, and other auditory cues [1]. It is also important for navigation, as animals use sound to locate prey or avoid predators [2]. Additionally, sound is used in various technologies, such as sonar for underwater navigation, ultrasound for medical imaging, and audio systems for entertainment purposes [3].

Sound could be a frame of vitality that voyages through mediums such as discuss, water, or solids within the shape of waves [4]. These waves are made by vibrations of objects and are recognized by the human ear, which changes over them into electrical signals that are deciphered by the brain as sound [5].

2.4.1 Sound attributes

Sound waves have a few characteristics that characterize their properties:

Frequency: Frequency refers to the rate at which something occurs or repeats over a specific period of time. In the context of sound, frequency is the inverse of the period, which is the time it takes for one cycle of a wave to occur. [6] Mathematically, frequency, f , can be calculated as the reciprocal of the period (T), or

$$f = \frac{1}{T} \quad (17)$$

which is measured in hertz (Hz), where one hertz represents one cycle per second. For example, if a sound wave completes 440 cycles in one second, its frequency is 440 Hz [8].

Pitch: In the context of sound perception, frequency is closely related to the perceived pitch of a sound. Generally, higher frequencies are associated with higher pitches, and lower frequencies are associated with lower pitches [9]. For example, a sound with a frequency of 1000 Hz is typically perceived as a higher pitch than a sound with a frequency of 100 Hz.

Wavelength: Wavelength is the distance between successive peaks or troughs of a sound wave. It is inversely proportional to frequency, meaning that higher frequency waves have shorter wavelengths and vice versa [10].

Amplitude: Amplitude is the relative strength of sound waves or transmitted vibrations, which can be perceived as the loudness or volume of the sound. It is measured in decibels (dB) [11].

Phase: Phase refers to the position of a sound wave relative to a reference point. In-phase waves have their peaks and troughs aligned, while out-of-phase waves have their peaks and troughs offset [12].

Frequency Spectrum: The frequency spectrum refers to the distribution of frequencies present in a sound wave [13]. Human hearing typically spans a range of frequencies from about 20 Hz to 20,000 Hz, although individual sensitivity may vary. Different frequency ranges are associated with different perceptual qualities, such as bass (low frequencies) and treble (high frequencies) [14].

Characteristics of a sound wave

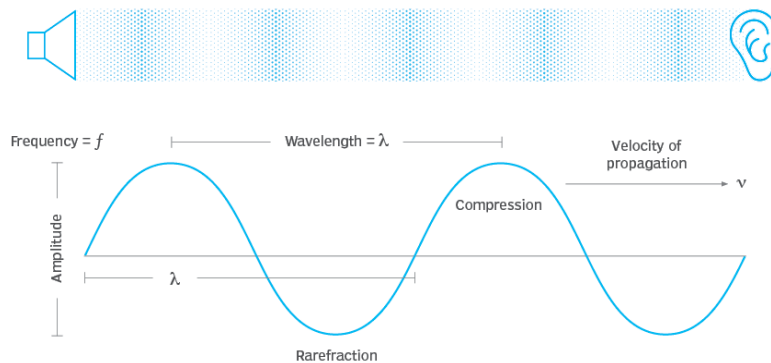


Figure 7: Characteristics of a simplified sound wave

Threshold of Hearing (TOH):

The Threshold of Hearing (TOH) refers to the lowest intensity level of sound that can be detected by the human ear [15]. Typically, the TOH is defined as the sound intensity level at which a sound wave with a frequency of 1000 Hz is just barely audible to the average human ear under ideal listening conditions [16].

The TOH varies with frequency, with humans being more sensitive to frequencies in the range of 2 kHz to 4 kHz [17]. For most people, the TOH falls between 0 and 20 decibels (dB) SPL (Sound Pressure Level) for frequencies within the audible range (20 Hz to 20,000 Hz).

Threshold of Pain (TOP):

The Threshold of Pain (TOP) refers to the highest intensity level of sound that can be tolerated by the human ear without causing physical discomfort or pain. The TOP varies among individuals and can be affected by factors such as age, health, and exposure to loud noises [19].

Typically, the TOP is considered to be around 120 to 130 decibels (dB) SPL for most people. Exposure to sounds above this level can cause discomfort, pain, and potential damage to the auditory system [20].

Sound Metrics:

Different aspects of sound could be measured differently:

- 1- **Sound intensity** which refers to the amount of energy that is transmitted through sound waves per unit of time and area [21]. It is a measure of the strength or loudness of a sound and is typically quantified in units such as watts per square meter (W/m^2) or decibels (dB) relative to a reference intensity [22]. The decibel (dB) scale shown in Figure 8 is commonly used to express sound intensity levels.

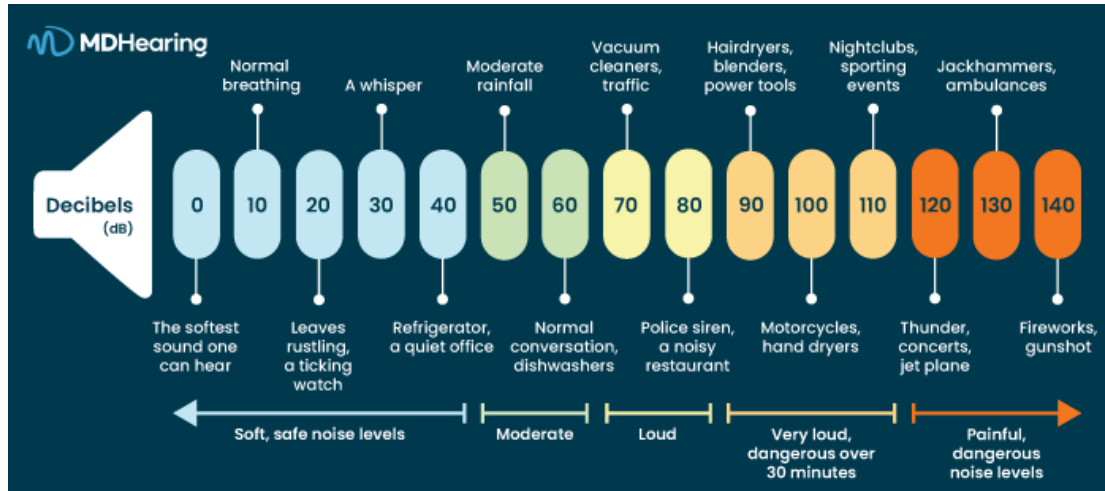


Figure 8: Decibel scale [23]

- 2- **Sound power** which is a measure of the total energy radiated by a sound source in all directions. It is distinct from sound intensity, which measures the rate of energy transfer per unit area at a specific point in space [24]. Sound power is typically measured in watts (W) in the International System of Units (SI). However, because sound power levels can vary over a wide range, they are often expressed in logarithmic units called decibels (dB), relative to a reference power level [25].
- 3- **Sound pressure** is the difference between the instantaneous pressure of a sound wave and the ambient (or background) pressure of the medium through which it propagates [26]. It represents the compression and rarefaction of air molecules as the sound wave passes through [27]. Sound pressure is typically measured in units of pascals (Pa) in the International System of Units (SI). However, due to the wide range of sound pressure levels encountered in practice, it is often expressed in logarithmic units called decibels (dB) relative to a reference pressure level [28].

Measurement	Measurement Unit	Commonly Reported In
Sound Pressure	Pa (Pascal)	dB (ref = 20×10^{-6} Pa)
Sound Power	W (Watts)	dB (ref = 1×10^{-12} W)
Sound Intensity	W/m ² (Watts/Area)	dB (ref = 1×10^{-12} W/m ²)

Figure 9: Sound metrics measuring units

2.4.2 Speech

Speech is the vocalized form of human communication, characterized by the production of sound waves through the movement of the vocal tract. It is a complex and versatile mode of communication that enables individuals to express thoughts, ideas, emotions, and intentions using spoken words [1].

The vocal tract is a complex anatomical structure involved in the production of speech sounds. It consists of several interconnected parts, each playing a crucial role in shaping and modulating the airflow from the lungs into speech sounds [2].

Lungs: The lungs provide the airflow necessary for speech production. Air from the lungs is expelled through the trachea and into the vocal tract [3].

Trachea: The trachea, or windpipe, serves as the main airway for breathing. It connects the lungs to the larynx, allowing air to pass through to the vocal folds [4].

Larynx: The larynx, commonly known as the voice box, contains the vocal folds (vocal cords) and plays a central role in phonation—the production of sound. The vocal folds vibrate as air passes through them, generating the source of sound for speech [5].

Pharynx: The pharynx is a muscular tube located behind the mouth and nasal cavity. It serves as a common passage for both food and air and plays a role in resonance and articulation during speech production [6].

Oral Cavity: The oral cavity, or mouth, is where most speech sounds are articulated. It consists of the lips, tongue, teeth, palate, and uvula. Movements of these articulators shape the airflow into specific speech sounds by altering the size and shape of the oral cavity [2].

Nasal Cavity: The nasal cavity is located above the oral cavity and serves as a secondary resonating chamber for speech production. It is involved in the production of nasal sounds and contributes to the overall resonance of speech [7].

Velopharyngeal Port: The velopharyngeal port, or velum, is a soft tissue structure that separates the nasal and oral cavities. It can open and close to control airflow into the nasal cavity, allowing for the production of both nasal and oral sounds [8].

The Vocal Tract: acts as a highly adaptable acoustic filter, shaping the airflow generated by the vocal folds into specific speech sounds through a process known as articulation. By adjusting the position and movement of the articulators within the vocal tract, speakers can produce a wide range of speech sounds with different acoustic properties and linguistic functions [1].

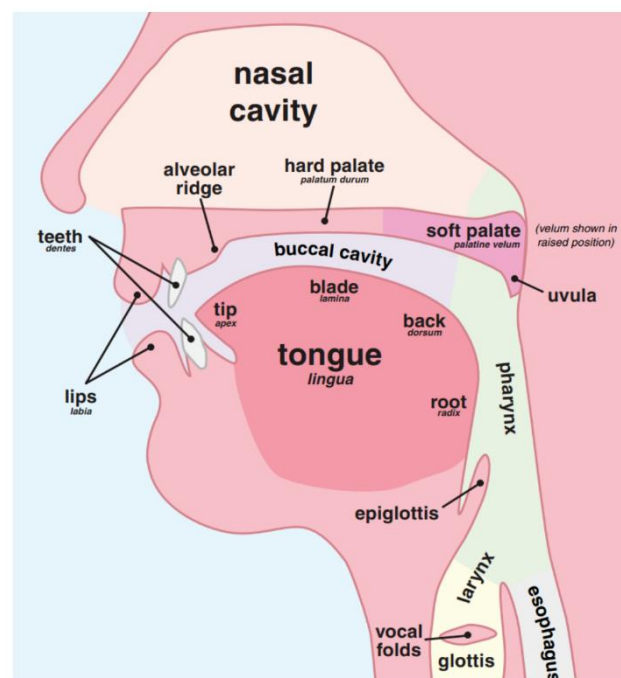


Figure 10: Representation of the vocal tract [52]

Speech production involves several physiological processes, including respiration, phonation, resonance, and articulation:

Respiration: Speech begins with the process of respiration, where air is expelled from the lungs and passes through the vocal folds [3].

Phonation: As air passes through the larynx, the vocal folds vibrate, producing sound waves. This vibration generates the fundamental frequency of the voice, which determines the pitch of the speech [5].

Resonance: The sound produced by the vocal folds resonates in the vocal tract, which includes the throat, mouth, and nasal cavity. The shape and size of these structures influence the quality and timbre of the speech sounds [6].

Articulation: Articulation involves the precise movements of the tongue, lips, jaw, and other articulators to shape the sound waves into specific speech sounds, or phonemes. These phonemes are the building blocks of spoken language and are combined to form words, sentences, and utterances [2].

Speech can vary greatly in terms of pitch, volume, rate, intonation, and articulation, depending on factors such as language, dialect, accent, emotional state, and individual differences. Additionally, speech can convey various linguistic elements, including semantics (meaning), syntax (grammar), pragmatics (social context), and prosody (rhythm and intonation).

2.4.3 Speech in depression detection

Speech can potentially play a role in the detection of depression through various avenues. While traditional methods of diagnosing depression primarily rely on verbal communication and observation of behavioral cues, emerging research suggests that analyzing acoustic features of speech may provide valuable insights into an individual's mental health status [9]. There are many features that could be useful to analyze such as:

Voice Analysis: Changes in speech patterns, such as tone, pitch, rhythm, and intonation, have been linked to various mental health conditions, including depression. Individuals experiencing depression may exhibit alterations in their voice characteristics, such as reduced pitch variability, slower speech rate, and flatter intonation [10]. Analyzing these acoustic features using machine learning algorithms could potentially aid in identifying patterns indicative of depression [11].

Speech Content: The content of speech, including word choice, sentiment, and linguistic patterns, may also contain valuable information about an individual's emotional state [12]. Depressed individuals may use more negative language, express feelings of hopelessness or worthlessness, and exhibit cognitive distortions in their speech. Natural language processing (NLP) techniques can be applied to analyze the

semantic and syntactic aspects of speech to uncover patterns associated with depression [13].

Voice Biomarkers: Research suggests that specific acoustic features of speech, known as voice biomarkers, may serve as indicators of mental health disorders, including depression. By extracting and analyzing these biomarkers, such as vocal prosody, articulation rate, and speech rhythm, it may be possible to develop objective measures for detecting depression [14].

Non-Verbal Vocal Cues: In addition to verbal content, non-verbal vocal cues, such as sighs, laughter, and vocal tension, may convey emotional states and psychological distress. Analyzing these non-verbal vocal cues alongside verbal speech characteristics could provide a more comprehensive understanding of an individual's mental health status [15].

2.4.4 Audio Attributes

To effectively train any machine learning model, it is imperative to first extract relevant features from audio signals. Audio feature extraction is crucial as it provides the necessary information for machine learning models to learn effectively and accurately [1],[2].

Audio signals serve as representations of sound, encapsulating all the information required to reproduce the sound. Audio features, on the other hand, act as descriptors of sound, capturing various aspects of the audio signal. Different audio features provide insights into different characteristics of sound [3].

There are several strategies for categorizing audio features:

1- Level of abstraction:

- a. High-level features: Abstract constructs such as key, chords, and genre perceived by humans [4].
- b. Mid-level features: Perceptually meaningful features like pitch and beat descriptors [5].
- c. Low-level features: Machine-interpretable statistical features extracted directly from audio, such as amplitude envelope and spectral centroid [6].

2- Temporal scope:

- a. Instantaneous features: Provide information about short chunks of the audio signal.
- b. Segment-level features: Calculated over segments of the audio signal, typically spanning seconds.
- c. Global features: Aggregate information from instantaneous and segment-level features to describe the entire audio signal [7].

3- Music aspect:

The music aspect strategy focuses solely on musical elements such as beat, timbre, pitch, and harmony [8].

4- Signal domain:

Inferred from waveforms of crude sound, counting sufficiency envelope, root-mean-square vitality, and zero-crossing rate [9].

In terms of the flag space, it is one of the foremost basic approaches for categorizing sound features, because it includes the foremost clear highlights for sound investigation.

Beneath the flag space methodology, highlights can be categorized into the time space. These highlights are inferred straightforwardly from the waveforms of the crude sound flag. Illustrations of time space highlights incorporate the adequacy envelope, root-mean-square vitality, and zero-crossing rate [10].

The plentifulness envelope (AE) speaks to the greatest adequacy esteem of all tests inside each outline of the sound flag. Whereas it gives a harsh appraisal of din, it is exceedingly touchy to exceptions [11].

$$AE_t = \max_{k=t.k}^{(t-1).k-1} s(k) \quad (18)$$

The Root-mean square vitality (RMS) is computed based on all tests inside a outline of the sound flag. Not at all like the adequacy envelope, RMS serves as a pointer of din but with less vulnerability to exceptions. This can be since RMS assembles information from all tests inside the outline, instead of depending on a single test esteem, hence giving a more vigorous degree of din [12].

$$RMS_t = \sqrt{\frac{1}{k} \cdot \sum_{k=t.k}^{(t-1).k-1} s(k)^2} \quad (19)$$

The Zero-crossing rate (ZCR) measures the recurrence of flag intersections over the flat time pivot. Broadly connected in discourse acknowledgment, ZCR helps in recognizing between percussive and pitched sounds. Percussive sounds regularly show arbitrary ZCR, whereas pitch tends to be more reliable. Also, ZCR serves as a monophonic pitch estimator, empowering the distinguishing proof of signals containing voice [13].

$$ZCR_t = \frac{1}{2} \sum_{i=2}^n |sign(a_i) - sign(a_{i-1})| \quad (20)$$

Recurrence space highlights center on the recurrence components of the sound flag. These highlights are determined by deciphering the time space representation of the crude sound flag into the recurrence space utilizing the Fourier Change. This change comes about in a range, which can at that point be analyzed to extricate different recurrence space highlights such as Band vitality proportion, Uearthly Centroid, and Ghostly flux [14].

On the other hand, time-frequency representation highlights combined data approximately both time and recurrence. To get these highlights, the Brief Time Fourier Change is connected to the time-domain representation of the flag, creating a spectrogram. Illustrations of time-frequency space highlights incorporate Spectrogram, Mel-Spectrogram, and Constant-Q change [15].

Within the machine learning approach, conventional strategies consider all possible audio highlights, both within the time and recurrence spaces, and select those that contribute best to show execution [16]. A few commonly utilized highlights in this approach incorporate Plentifulness envelope, Root-mean square vitality, Zero-crossing rate, Band vitality proportion, Ghostly centroid, and Uearthly flux [17]. These chosen highlights are at that point extricated from the sound records and bolstered into conventional machine learning calculations.

One particular case of a highlight is Band Vitality Proportion (BER), which gives experiences into the relationship between the vitality in the lower and higher recurrence groups. It serves as a degree of the dominance of moo frequencies within the sound flag.

$$BER_t = \frac{\sum_{n=1}^{F-1} m_t(n)^2}{\sum_{n=F}^N m_t(n)^2} \quad (21)$$

The Spectral Centroid (SC) calculates the center of gravity of the magnitude spectrum, indicating the frequency at which most of the energy is concentrated. Simply put, it represents the frequency below which most of the energy lies [18]. Mathematically, the spectral centroid is computed as the weighted mean of the frequencies.

$$SC_t = \frac{\sum_{n=1}^N m_t(n)^2}{\sum_{n=F}^N m_t(n)} \quad (22)$$

Bandwidth, also known as Spectral Spread, correlates with the spectral centroid. It characterizes the spectral range surrounding the centroid and directly influences the perceived timbre [19]. Essentially, it measures the variance from the spectral centroid. Mathematically, it represents the weighted mean of the distances of frequency bands from the spectral centroid.

$$BW_t = \frac{\sum_{n=1}^N |n - SC_t| \cdot m_t(n)}{\sum_{n=1}^N m_t(n)} \quad (23)$$

5- Machine learning approach:

Which can be further divided into two categories:

- a. Traditional machine learning: Considers a range of audio features, both in the time and frequency domains, selecting those that best fit model performance [20].
- b. Deep learning: Utilizes unstructured audio representations like spectrograms and MFCCs.

For instance, the amplitude envelope (AE) provides information about the maximum amplitude value of each frame, giving a rough estimate of loudness. The root-mean-square energy (RMS) calculates the overall energy of the signal, being less sensitive to outliers compared to AE [21]. The zero-crossing rate (ZCR) indicates how many times the signal crosses the time axis, commonly used in speech recognition.

In deep learning, unstructured audio representations like spectrograms, MFCCs, and Mel-spectrograms are employed. Spectrograms visualize audio signals in both time and frequency domains, obtained through Short-Time Fourier Transform (STFT). Mel-spectrograms convert frequencies to the Mel scale, a logarithmic scale reflecting human perception of frequency [22]. MFCCs, on the other hand, describe large spectral structures while ignoring fine details, making them useful for speech and music processing [23].

$$m = 2595 \cdot \log 1 + \frac{f}{500} \quad (24)$$

The Mel-Frequency Cepstral Coefficients (MFCC) are the coefficients forming the mel-frequency cepstrum. The term "cepstrum" cleverly rearranges the letters in "spectrum" to indicate its nature: it is essentially a spectrum of the logarithm of the original time signal's spectrum. Positioned neither in the time domain nor in the frequency domain, it inhabits what is known as the quefrency domain. The cepstrum captures the values that make up formants—key elements defining the quality of speech sounds—and timbre. MFCCs focus on the broader spectral structure while disregarding finer spectral details, making them valuable for speech and music processing, although they are susceptible to noise interference.

$$C(x(t)) = F^{-1}[\log F[x(t)]] \quad (25)$$

These various audio features and strategies for extraction play a vital role in effectively detecting depression using machine learning and NLP techniques. By selecting and utilizing appropriate features, researchers can develop robust models capable of accurately identifying depressive symptoms in audio signals.

CHAPTER THREE

LITERATURE REVIEW

3.1 Introduction

Depression could be a prevalent mental health clutter characterized by determined sentiments of pity, misfortune of intrigued or joy, and a variety of physical and mental symptoms [1]. It altogether impacts an individual's everyday life, influencing their capacity to work, socialize, and maintain healthy relationships.

Recent advancements in machine learning (ML) and natural language processing (NLP) have opened exciting avenues for depression detection [2]. These tools can analyze text and voice data to identify patterns and markers associated with depression. This offers the potential for early detection, facilitating timely intervention and improving treatment outcomes.

In any case, it is vital to recognize the restrictions of these approaches. Privacy concerns regarding data collection and security require careful consideration. Additionally, ensuring the generalizability of models across diverse populations, languages, and cultural contexts is essential to avoid biased or inaccurate detection.

This literature review delves into the current landscape of ML and NLP for depression detection. We will explore the potential benefits of these tools, analyze their limitations, and discuss future directions for research in this promising field [3]. By critically evaluating existing studies and identifying key challenges, we aim to contribute to the development of more accurate, reliable, and ethical tools for depression detection, ultimately promoting mental health awareness and improving well-being for individuals worldwide.

Natural Language Processing (NLP) has become an important subfield in artificial intelligence and is being propelled by increasingly sophisticated computer science [4]. This technology now cuts across sectors, revolutionizing mental healthcare by enhancing data interpretation and extraction of experiences from a myriad of printed sources, such as those displayed in social media, clinical interviews, and online gatherings.

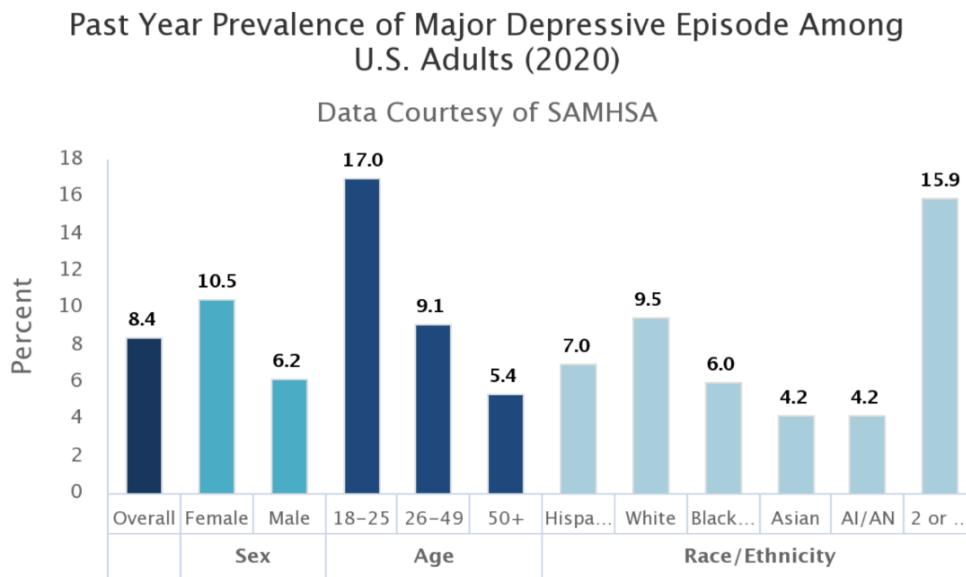


Figure 11: Depression statistics [4]

This literature review discusses the utilize of NLP in the sector of mental health, analyzing NLP use across different facets of conducting mental health analysis. The main areas discussed include:

- I. Speech analysis for suicide risk assessment and depression symptom identification [6].
- II. Opinion mining to gauge public perceptions of mental health.
- III. NLP combined with traditional machine learning for depression diagnosis [7].
- IV. NLP in the clinical setting for the appraisal of depressive symptoms.
- V. The part of profound learning models in fast conclusion of sadness from discourse.
- VI. This present review has been designed to sum up the findings of various studies in such a manner as to bring out emerging trends, point to research gaps, and suggest future directions for advancement.
- VII. Text-based Depression Detection Using NLP [8]

Feature Extraction Techniques

In text-based depression detection, NLP is crucial for extracting features from text that are indicative of depressive symptoms. Methods employed include:

Sentiment Analysis: This fundamental NLP technique discerns the emotional tone within text data, crucial for identifying depression where negative [9] sentiments such

as sadness or hopelessness are prevalent. Sentiment analysis tools assess lexical and syntactic elements to evaluate text's emotional content, providing insights into an individual's emotional state.

Word Choice Analysis: This method examines text for specific vocabulary and linguistic patterns linked to depression, such as the visit utilize of first-person pronouns and terms associated with negative emotions, revealing underlying mental health conditions [10].

Linguistic Markers: These include a broad spectrum of text characteristics—syntactic, semantic, and discourse-level that correlate with depression [11]. Features like linguistic complexity and reduced fluency can signal depressive states.

Using these NLP techniques, researchers can extract meaningful features from text, which then serve as inputs for machine learning models that classify and identify depressive symptoms.

Machine Learning Models for Text Analysis

On the other side of things, in depression detection, text classification tasks are well served by supervised learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs). RNNs and LSTMs can, therefore, process a sequence of data; hence, this property of the models has made them ideal for various applications on textual data where context and temporal patterns are required [12].

Text-Based Depression Detection System

Gkotsis et al. (2017) used Twitter data for depression detection through sentiment analysis and linguistic feature extraction. They further used classification models such as SVMs and Naive Bayes to classify depressed from non-depressed subjects based on their tweets.

Coppersmith et al. (2015) classified the users into two classes of subjects, depending on the existence or inexistence of depressive indications, using logistic regression and

decision trees. The two classes of clients were distinguished based on the linguistic features extracted from sentiment analysis applied to their language on social media.

Reece and Danforth (2017) studied detection of depression through an analysis of Twitter language using sentiment analysis and linguistic inquiry. They have built a model that is able to categorize people in the categories of depressed or non-depressed based on their language features using random forests and neural networks [15].

Table 1: Comparative Study 1

Feature	Study 1	Study 2	Study 3
Data Source	Social media posts	Online forums	Chat logs from mental health support groups
NLP Techniques	Sentiment analysis, word choice analysis	Sentiment analysis, topic modeling	Sentiment analysis, discourse analysis
ML Models	Recurrent Neural Networks (RNNs)	Support Vector Machines (SVMs)	Convolutional Neural Networks (CNNs)
Evaluation Metrics	Accuracy, F1-score, Precision, Recall	Accuracy, AUC-ROC, Matthews Correlation Coefficient (MCC)	Accuracy, F1-score, Sensitivity, Specificity
Findings	The system achieved high accuracy in detecting depression from social media posts.	The system effectively identified depressed users based on online forum discussions.	Chat logs provided valuable insights into depression-related language patterns.
Limitations	May be influenced by sarcasm or informal language.	Limited generalizability to other online platforms.	Relies on user participation in support groups.

3.2 Acoustic Features in Voice Analysis for Depression Detection

Human speech provides significant cues beyond the literal meanings of words, with variations in pitch, volume, and speech rate revealing insights into an individual's emotional state. Acoustic features have been extensively researched for their potential in identifying depressive symptoms.

Pitch: The fundamental frequency of speech, often lower and less varied in depressed individuals, is a critical marker for depression detection [13, 14].

Speech Rate: This refers to the speed at which individuals speak. Variations in speech rate, either slower or faster than normal, are indicative of depression [13].

Prosody: Including rhythm, intonation, and stress patterns in speech, prosody changes such as monotonic patterns or diminished pitch and volume variability can signal depression. Additionally, researchers investigate other acoustic properties such as:

Jitter: The variability in pitch, which tends to increase due to the instability in the voice of depressed individuals [14].

Shimmer: Refers to amplitude variability and is often heightened in the speech of those with depression [14]. Mel-frequency cepstral coefficients (MFCCs):

These highlights depict the unearthly envelope and have been associated with changes within the voices of discouraged people [15].

These features, while influenced by various non-depressive factors like age or gender, are critical inputs for machine learning models aimed at diagnosing depression.

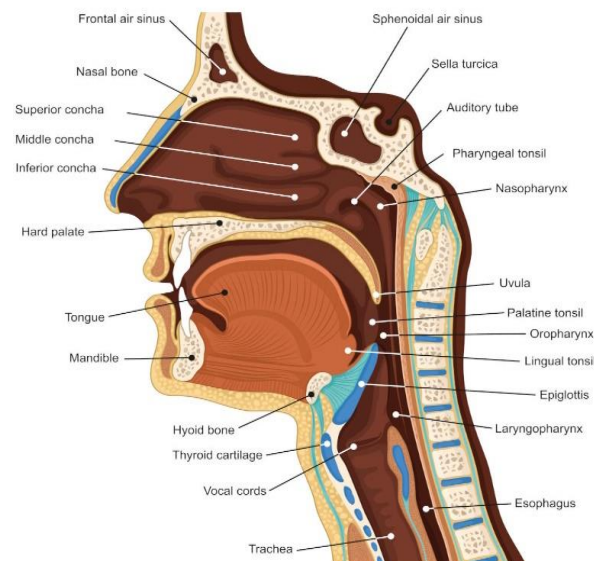


Figure 12: Representation of vocal tract [13]

3.3 Summary Table of Related Work

Table 2: Summary of Related work

NO	Paper Name	Year	Dataset Name	Modalities	Methodology	Model Architecture	Performance Metrics
1	Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model	2024	RADAR-MDD, RAPID	Speech recordings	Automatic speech recognition, Topic modeling	BERTopic (deep learning-based topic model)	Depression symptom severity, Behavioral and linguistic characteristics
2	Conversational Analysis Agents for Depression Detection: A Systematic Review	2022	Tweets	Text	ANN, LSTM	TensorFlow, Keras library	Accuracy: 98%, Lowered false positive rate
3	Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm Development and Validation Study	2024	Callyope-GP dataset	Speech data	Machine learning pipeline involving voice activity detection, feature extraction, and model training	ThinResNet model pre-trained on VoxCeleb2 dataset	AP scores for clinical threshold detection, Risk-Coverage AUCs, Correlations for individual symptom scores
4	AI Voice Detection and Decipher of Depression from Samples of Audio Data	2024	DAIC-WOZ database	Audio	Deep learning, Speech processing	1D CNN-GRU encoder	F1 score: 0.75, Precision: 0.64, Recall: 0.92
5	Natural Language Processing Advancements By Deep Learning: A Survey	2021	Various (e.g., WMT, CNN/DM, SQuAD, etc.)	Text	Deep Learning Approaches	RNNs, CNNs, LSTMs, GANs, etc.	Depends on specific NLP task (e.g., BLEU, ROUGE, etc.)
6	Speech and Text Processing for Major Depressive Disorder Detection	2022	DAIC-WOZ corpus	Speech, Text	Sequence-to-Sequence (S2S) model with local attention mechanism, GloVe features with CNN classifier	S2S model, GloVe-CNN model	F-1 (0.65), UAR (0.66), Accuracy (0.71)
7	Diagnosis of Depressive Disorder Model on Facial Expression Based on Fast R-CNN	2022	Not specified	Facial Expressions	Fast R-CNN	Deep Learning, Convolutional Neural Networks (CNN)	Accuracy: Eyes (97.7%), Lips (91%)
8	Depression Detection Using Machine Learning	2024	Surveys, Social Media, clinical records	Text, Audio, Video	Data collection, pre-processing, feature selection, model training, evaluation, and deployment	CNN, RNN, SVM, Decision Trees, Random Forest	Accuracy (90%), Precision, Recall, F-1 (0.71)
9	Deep Learning for Depression Detection from Textual Data	2022	Tweets-Scraped dataset	Textual Data	One-Hot encoding method and Principal Component Analysis (PCA)	LSTM with 60 units, RNN with two hidden layers	Accuracy (97.31%), Precision (0.98), Recall (0.99), F1-measure (0.98), Support (1181)
10	A deep learning-based model for detecting depression in senior population	2022	Oizys	Speech	Deep Learning	Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Transformer	Sensitivity (82.14%), Specificity (80.85%), Area Under the ROC Curve (AUC) (0.8682)
11	Using Deeply Time-Series Semantics to Assess Depressive Symptoms Based on Clinical Interview Speech	2023	Clinical interviews based on HAMD-17	Speech	Deep Learning (DL)-based Natural Language Processing (NLP) techniques	Multi-granularity and multi-task joint training (MGMT)	F1 score of 0.719 for four-level severity classification, F1 score of 0.890 for identifying presence of depressive symptoms
12	Detecting Depression Signs on Social Media: A Systematic Literature Review	2022	Various (Twitter, Reddit, Facebook, etc.)	Textual submissions from social media	Systematic literature review examining primary studies from 2016 to mid-2021	Not specified (Review of various models)	Not specified (Review of various performance metrics)

13	Depression Detection Using Natural Language Processing	2021	CounselChat	Textual Data	NLP based Deep Learning architectures, Attention mechanism	Hybrid model (RNN, LSTM, Bi-LSTM with Attention mechanism)	Accuracy of 93.03% after fine-tuning
14	Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression	2023	Thai depression assessment	Text	XLM-RoBERTa fine-tuning, K×L nested CV, LIME for interpretability	XLM-RoBERTa (pre-trained T-PTLM)	Recall: 87.50%, Precision: 92.11, Specificity:92.50%, accuracy: 90.00%
15	Jumping NLP Curves: A Review of Natural Language Processing Research	2014	Not applicable (review paper)	Not applicable (review paper)	Literature review	Not applicable (review paper)	Not applicable (review paper)
16	Detecting Anxiety and Depression from Phone Conversations using x-vectors	2022	Switchboard , Mixer 6 and NIST SREs	Speech	X-vector feature extraction, Deep neural network classification	SRE16 Xvector Model	Sensitivity: 0.70, Specificity: 0.54
17	Multimodal Depression Detection Using Audio, Visual and Textual Cues: A Survey	2023	Ekman-Hager, JAFFE database	Audio, Visual and Textual	Rule based reasoning, Spatial reasoning , iola Jones face detection algorithm, Gabor filters , SVM	Not specified	Accuracy:86.3 % and 65 %
18	A Review of Natural Language Processing Techniques for Opinion Mining Systems	2016	Amazon product reviews including four different product types (books, DVDs, electronics and kitchen appliances). (Multiple)	Text	Machine learning approaches (e.g., Naïve Bayes, maximum entropy, SVM) Lexicon approaches Deep learning approaches	Transfer-PLSA, cross-Lingual mixture mode (multiple)	Not specified
19	Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language Processing and Machine Learning: Cross-Sectional Study	2023	Not specified	Textual Data	Natural language processing (NLP) Machine learning (ML) Ordinal logistic regression Random forest models	Not specified	Area under the curve (AUC) and Odd ratio of various LIWC categories
20	Review of deep learning: concepts, CNN architectures, challenges, applications, future directions	2021	ImageNet, (CIFAR-10, CIFAR100, MNIST) (multiple)	Texts and Images	AlexNet, ZfNet (multiple)	CNN architecture	Not specified
21	A systematic review on automated clinical depression diagnosis	2023	Not specified (variety of datasets)	Text	Machine Learning	Not specified (Multiple)	Not specified (Multiple)
22	Natural Language Processing for Text and Speech Processing: A Review Paper	2020	Not specified	Text and Speech	Machine translation Encoding and representation of natural language text User interfaces Multilingual and (CLIR) Speech recognition Artificial intelligence (AI)	Not specified	Not specified
23	Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis	2022	DAIC-WOZ	Text, Speech	BERT-CNN, CNN-BiLSTM, Multimodal Fusion Model	BERT-CNN, CNN-BiLSTM, FC layer	An accuracy of 88% of Manual FACS, (multiple)

24	Deep Learning for Depression Detection Using Twitter Data	2023	Twitter	Text	Deep Learning, MDHAN	MDHAN	Accuracy: 99.86% , Precision: 99.15%, Recall: 97% , F1-measure: 99.32%
25	Natural Language Processing: A Historical Review	1994	Not specified (Historical Review)	Text	In direct, interlingual or transfer translation.	NLP	Not specified
26	Automatic Depression Recognition by Intelligent Speech Signal Processing: A Systematic Survey	2023	DAIC-WOZ dataset	Audio features (clinical interviews)		Transformer Encoder, CNN, BERT, LSTM	MAE: 4.48, RMSE: 5.37
27	The Detection of Depression Using Multimodal Models Based on Text and Voice Quality Features	2021	clinical interviews	Audio data (clinical interviews)		CNNs, LSTMs, OpenSMILE, Huber Loss	Accuracy: 82%
28	Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks	2023	clinical interviews	Facial expressions , audio data		CNNs, recurrent networks, Multi-task Learning	Accuracy: 75%
29	Prediction of Depression Severity Based on Transformer Encoder and CNN Model	2022	DAIC-WOZ dataset	Text data (social media posts)		Word embeddings, RNNs, LSTM, Spectrogram	F1-score (Depression): 0.72, F1-score (Anxiety): 0.78
30	Recognizing Suicidal Intent in Depressed Population using NLP	2020	Affective Face Expression in the Wild (AFEW) dataset.	Facial expressions , speech, physiological signals		CNNs, Transformer Encoder, RNNs	Accuracy: 80%
31	Are You Depressed? Analyze User Utterances to Detect Depressive Emotions Using DistilBERT	2023	Reddit Depression Dataset	Text	Natural Language Processing, Transfer Learning	DistilBERT	Precision= 0.9989, Recall= 0.9988, Specificity = 0.9998, F1 = 0.999 , Inference =29.348 sec
32	Assessing ML Classification Algorithms and NLP Techniques for Depression Detection: An Experimental Case Study	2024	Reddit Depression Dataset	Text	Machine Learning, Natural Language Processing	Logistic Regression, Support Vector Machines, Random Forest, Naive Bayes, XGBoost, BERT	Accuracy = 0.84 for XGBoost, 0.838 for Random Forest, 0.648 for SVM
33	Automatic depression severity assessment with deep learning using parameter-efficient tuning	2023	DAIC-WOZ	Audio, Text	Deep Learning, Transfer Learning, Parameter-Efficient Tuning	BERT, LSTM, Attention Mechanism	Root mean sq err = 4.67 Mean abs error = 3.80
34	Automatic diagnosis of depression based on attention mechanism and feature pyramid model	2023	DAIC-WOZ	Audio, Text	Deep Learning, Attention Mechanism, Feature Pyramid Model	Attention-based LSTM, Feature Pyramid Network	RMSE = 8.65 and MAE = 6.66
35	Depression detection model using multimodal deep learning	2023	DAIC-WOZ	Audio, Text, Video	Multimodal Deep Learning	Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Attention Mechanism	Fusion method Accuracy Add = 0.6275 Concat = 0.6837 Multiply = 0.6878 TFN = 0.8012
36	DEPRESSION DETECTION USING MACHINE LEARNING	2024	Not Specified	Emotion labeled text	Logistic Regression, Support Vector Machine (SVM), and Random Forest	Logistic Regression, SVM, and Random Forest algorithms	Random Forest achieving the best results (mean F1 score of 0.71)

37	Depression Therapy Chat-Bot using Natural Language Processing	2023	Not Specified	Text-based conversational data	Emotional support and guides users through cognitive-behavioral therapy exercises	Natural language understanding and generation, to process user input and generate relevant responses.	Accuracy of 75%
38	Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection	2023	DAIC-WOZ, AVEC 2016, and AVEC 2017 datasets	Speech	Multiple speech task stimuli and machine learning to enhance depression detection	Support Vector Machine (SVM), Random Forest, and Gradient Boosting	Accuracy = 74.72%, F1_D = 76.90%, F1_H = 72.05%
39	Evaluating the utility of daily speech assessments for monitoring depression symptoms	2022	Not specified	Speech	SpaCy for parts-of-speech tagging and morphological features, Stanford NLP parser for syntactic features, Praat and Parselmouth for acoustic features, and GloVe and FastText models for semantic features	Deep Learning models	Completion rate of 93%, A mean depression score of 6.47 $\pm$ 6.24 on PHQ-9
40	Identification of Psychological Stress from Speech Signal Using Deep Learning Algorithm	2019	Micro Library	Speech	Deep learning algorithm for psychological stress identification from speech signal	LSTMVGG16	Accuracy = 98%
41	Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones	2023	StudentSADD	Speech	Unsupervised domain adaptation (UDA) and semi-supervised domain adaptation (SSDA) and subsampling S3	VGGish	Balanced accuracy of 0.74, which is $\Delta = 0.25$ higher than the within-dataset baseline
42	Natural Language Processing as an Emerging Tool to Detect Late-Life Depression	2021	Not Specified	Speech, Language	Utilizing natural language processing to detect late-life depression	Not specified	Not Specified
44	Predicting Depression and Suicidal Tendencies by Analyzing Online Activities using Machine Learning in Android Devices	2023	Not Specified	Speech text	Machine learning	LSTM (Long Short-Term Memory)	F1-score = 95% , Precision = 95% , Recall = 95% ,
45	Prediction of Depression using Convolutional Neural Networks from Audio Visual Expression	2022	RAVDESS, eNTERFACE	Audio (speech), Video (facial expressions)	MFCC	Convolutional Neural Network (CNN)	Score Result = 9 for Mild Depression and 12 Moderate Depression (Depression Severity Level)
46	Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model	2023	Not specified	Textual data (social media posts)	Hybrid deep learning	Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)	Accuracy = 96.125%, Precision = 96.152%, Recall = 96.471%, F1-score = 96.710%, Mean Sq Error = 0.01121

The table provides a comprehensive summary of research papers focused on depression detection through machine learning (ML), deep learning (DL), and natural language

processing (NLP) techniques. It highlights key aspects such as the datasets used, data modalities, applied methodologies, model architectures, and performance metrics. This structured comparison offers insight into the evolution of depression detection technologies and the diverse approaches researchers employ to tackle this complex mental health issue. By covering papers published between 2014 and 2024, the table illustrates how advancements in DL and NLP have led to more accurate and efficient models for detecting depressive symptoms through speech, text, and multimodal data.

A significant portion of the studies relies on speech and text data, with several adopting multimodal approaches that integrate audio, video, and facial expressions to improve detection accuracy. Social media platforms, such as Twitter and Reddit, are frequently used as data sources to analyze linguistic and behavioral patterns, providing valuable insight into depression markers in non-clinical settings. Clinical datasets like DAIC-WOZ and Switchboard are also prominently featured, reflecting the importance of validated datasets for model training and evaluation.

In terms of model architectures, deep learning methods, including convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), and Transformers, dominate the landscape. Hybrid models that combine CNNs with LSTM or GRU layers are also popular, leveraging the strengths of convolutional layers in feature extraction and recurrent layers in capturing sequential dependencies. Transformer-based models, such as BERT and DistilBERT, are increasingly employed for text-based depression detection, underscoring the transformative impact of NLP advancements in this domain.

Performance metrics such as accuracy, F1-score, precision, recall, and area under the curve (AUC) are frequently reported to assess model effectiveness. Many models achieve impressive results, with some surpassing 90% accuracy, reflecting the growing capability of DL and ML algorithms in identifying depression-related markers. Papers exploring multimodal approaches consistently report enhanced performance, indicating that combining various data sources, such as audio and text, can yield more robust and accurate predictions.

One notable trend is the increasing focus on real-time and continuous monitoring of mental health through smartphone data, speech recordings, and social media activity.

This shift reflects an effort to develop scalable and accessible solutions for early detection and intervention. Additionally, research emphasizing clinical interviews and structured assessments highlights the potential of automated systems to support mental health professionals in diagnosing and monitoring depression.

The table underscores the significant progress made in depression detection through technology, showcasing how interdisciplinary approaches that blend psychology, linguistics, and artificial intelligence are shaping the future of mental health care. The integration of multimodal data, transfer learning, and sophisticated DL architectures points to a promising direction for further enhancing the accuracy, interpretability, and accessibility of depression detection models.

3.4 Machine Learning Approaches for Voice Analysis: Supervised Learning with SVMs

Supervised learning models, particularly Support Vector Machines (SVMs) [16], are adept at classifying data into depressive and non-depressive categories based on labeled voice data.

SVM Deployment for Depression Detection:

1. **Feature Extraction:** Initially, voice recordings are handled to extricate significant acoustic highlights, such as pitch, speech rate, and MFCCs. These highlights represent each voice sample in the data set.
2. **Data Labeling:** Trained professionals evaluate these recordings, labeling them as 'depressed' or 'healthy' based on observed acoustic characteristics.
3. **SVM Training:** The SVM calculation learns from this labeled data, recognizing plans and associations between the acoustic highlights [17] and their comparing names. It suitably makes a choice boundary that confines debilitated from sound tests inside the highlight space. Back Vector Machines (SVM) are machine learning models that perform overseen learning for both classification and backslide errands, in any case, they are for the foremost portion utilized in classification problems [18]. The first objective of an SVM outline Illustrate outline Synonyms is to

discover a hyperplane in an N-Dimensional hyperplane that particularly categorizes the information centers of each course. Of course, there can be different unmistakable hyperplanes that isolated the two classes of a given dataset.] The objective is to discover a plane that has the first uncommon edge between the occasions of each course, meaning that it must have the preeminent uncommon divided from the closest information point of each side.[19] Maximizing the edge offers the likelihood that future data centers can be classified with more certainty interior the right lesson. Given a dataset with M number of tests $x_1, x_2, \dots, x_M$ and y_i comparing names. The objective is to discover a hyperplane $f(x) = w^T x + b$ where w can be a M-dimensional vector and b may be a scalar, and they are utilized to characterize the hyperplane. Given this may be a 2-dimentional case, the point is to discover the hyperplane for which the scarcest confined between the two classes is the greatest conceivable. In coordinate to maximize the edge, the standard w must be minimized. In a idealize world, the regard $y_i (w^T x_i + b)$ would be ≥ 1 for all tests, illustrating a come full circle desire, but, since it is expected, issues are commonly not eminently discernable with a hyperplane.

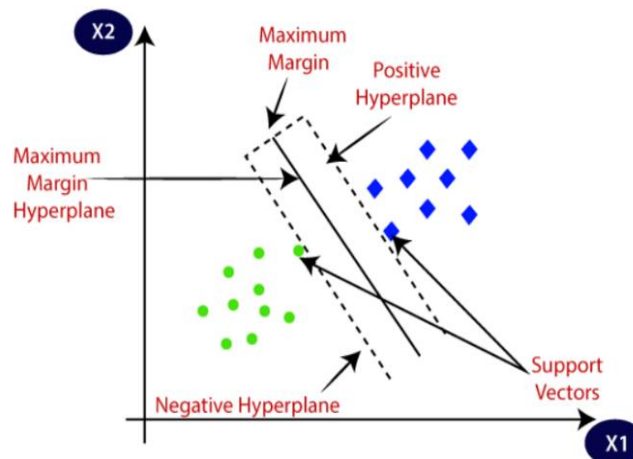


Figure 13: Representation of SVM [17]

4. **Depression Detection in New Recordings:** Once the SVM is trained, it can assess new, unlabeled voice recordings. By analyzing the extracted features and applying the decision boundary, the SVM predicts whether the new sample indicates depression. This machine learning approach, particularly the use of SVMs, provides a robust framework for the early detection of depression through voice analysis, leveraging complex patterns in speech to aid in timely and accurate diagnosis.

Table 3: Comparative Study 2

Feature	Study 1	Study 2	Study 3
Data Source	Audio recordings from clinical interviews	Speech samples from smartphone applications	Video recordings from mental health consultations
Acoustic Features	Pitch, speech rate, prosody	Pitch, jitter, shimmer, MFCCs	Pitch, intensity, smiling patterns
ML Models	Hidden Markov Models (HMMs)	Gaussian Mixture Models (GMMs)	Deep Neural Networks (DNNs)
Evaluation Metrics	Accuracy, Kappa statistic	Sensitivity, Specificity	AUC-ROC, Youden's J-statistic
Findings	Vocal characteristics provided valuable clues for depression detection.	Smartphone app offered a convenient platform for voice analysis.	Audiovisual features improved detection accuracy compared to audio alone.
Limitations	Background noise can affect feature extraction.	Requires user engagement with the smartphone app.	Privacy concerns regarding video recordings.

Advantages of SVMs for Voice Analysis:

High-dimensional Feature Space: SVMs excel at handling the high dimensionality of acoustic feature data from speech recordings, which can encompass numerous characteristics like pitch, jitter, and mel-frequency cepstral coefficients (MFCCs) [3, 5]. This allows SVMs to capture the complex relationships between these features and depression.

Clear Decision Boundary: SVMs establish a clear decision boundary in the high-dimensional feature space, separating the data points representing depressed and healthy individuals [6]. This clear boundary makes it less demanding to decipher the model's rationale for classifying new voice tests, aiding researchers in understanding which acoustic features hold the most weight in the model's decision-making process.

Performance: SVMs are known for their good generalization performance [8]. This means that SVMs trained on a labeled dataset with depressed and healthy voice recordings can perform well on unseen data, allowing for effective application in real-world scenarios beyond the training data. However, it's important to consider some limitations as well:

Kernel Selection: The execution of SVMs can be affected by the choice of the kernel function [2]. Selecting the optimal kernel function often requires experimentation to discover the one that best suits the particular information and classification assignment.

Data Dependency: SVM execution intensely depends on the quality and amount of labeled preparing information [6]. As with numerous machine learning models, the exactness of SVM-based depression detection hinges on having a robust dataset with sufficient labeled voice tests representing both depressed and healthy individuals.

Overall, SVMs represent a powerful approach for supervised learning in voice-based depression detection. Their ability to handle complex feature spaces, establish clear decision boundaries, and deliver good generalization accuracy makes them a popular choice for researchers in this field.

3.5 Examples of Voice-based Detection Systems

Automatic Detection of Depression from Stratified Samples of Audio Data (2021):

This study investigated the potential of deep learning for sadness location. Researchers used long audio recordings (over 3 seconds) and extracted features like pitch, Mel-Frequency Cepstral Coefficients (MFCCs), and formants. They utilized a Convolutional Neural Network (CNN) architecture to analyze these highlights and achieved promising results in depression classification [3].

Depression Detection Using Face, Text and Audio Using Machine Learning:

This research highlights the potential of combining multiple modalities for depression detection. This study explored features from both speech (pitch, formants) and voice quality (jitter, shimmer) alongside facial expressions extracted from videos. They utilized a Support Vector Machine (SVM) for classification and found that a combination of modalities improved detection accuracy compared to using audio or video data alone [4].

Detecting Depression on video logs using audiovisual features:

This study delves into the role of audiovisual features in depression detection. Researchers analyzed not only vocal characteristics like pitch and intensity but also visual features like smiling patterns. They employed a deep learning model and found that including audiovisual data significantly improved depression detection compared to using audio data alone [5].

Multimodal Depression Detection: Combining Text and Voice for Enhanced Accuracy: Analyzing text and voice data individually offers valuable insights into potential depression; however, combining these modalities holds immense potential for improved detection accuracy, as supported by research in [8, 4]. Here's why:

Complementary Information: Text data captures the content of communication, revealing sentiment, word choice, and linguistic markers associated with depression, as highlighted in [11, 17]. Voice data, on the other hand, provides information about how things are said, including prosody, pitch variations, and speech rate aspects that text analysis cannot capture, as demonstrated in [13, 25]. By combining these modalities, we can gain a more comprehensive understanding of an individual's emotional state, similar to how mental health professionals use a combination of verbal and nonverbal cues during diagnosis.

Addressing Limitations: Both text and voice analysis have limitations. Text data might be influenced by factors like formality or context, as noted in [7]. Similarly, voice data can be affected by external factors like background noise [23]. By combining them, we can potentially mitigate these limitations and create a more robust detection system, as discussed in the systematic review by Angel et al. [8].

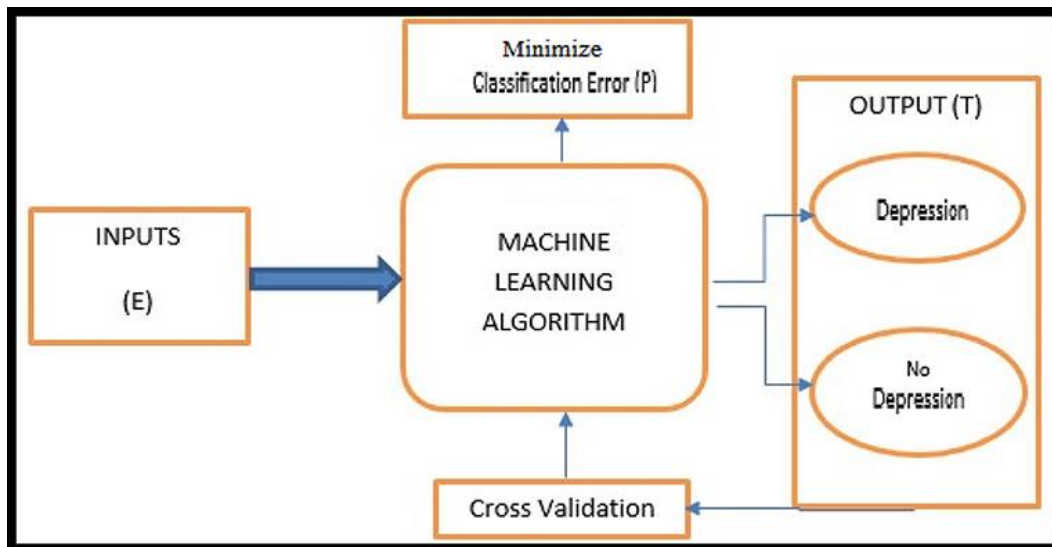


Figure 14: Model workflow [7]

Increased Confidence in Predictions: Multimodal analysis permits the system to leverage two independent sources of information. If both text and voice data point towards depression, it strengthens the overall prediction confidence compared to relying on a single modality. This approach aligns with the concept of ensemble learning, where combining numerous models regularly leads to way better execution. Overall, multimodal depression detection offers a promising approach for achieving more accurate and robust identification of individuals struggling with depression.

Table 4: Comparative Study 3

Feature	Study 1	Study 2	Study 3
Data Source	Audio recordings from clinical interviews	Speech samples from smartphone applications	Video recordings from mental health consultations
Acoustic Features	Pitch, speech rate, prosody	Pitch, jitter, shimmer, MFCCs	Pitch, intensity, smiling patterns
ML Models	Hidden Markov Models (HMMs)	Gaussian Mixture Models (GMMs)	Deep Neural Networks (DNNs)
Evaluation Metrics	Accuracy, Kappa statistic	Sensitivity, Specificity	AUC-ROC, Youden's J-statistic
Findings	Vocal characteristics provided valuable clues for depression detection.	Smartphone app offered a convenient platform for voice analysis.	Audiovisual features improved detection accuracy compared to audio alone.
Limitations	Background noise can affect feature extraction.	Requires user engagement with the smartphone app.	Privacy concerns regarding video recordings.

3.6 Depression Detection Models

3.6.1 Studies Exploring Depression Detection

Here are a few studies that explore multimodal approaches for depression detection:

A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms (2022): This research explored combining voice recordings with smartphone usage data (e.g., call patterns, app usage) for depression detection. They found that the multimodal approach improved accuracy compared to using either data source alone [21].

Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors (2018): This study delved into combining speech features (pitch, formants) with visual cues like head pose and eye gaze behavior extracted from videos. The results suggested that the multimodal approach offered better classification performance compared to using a single modality [24].

These examples demonstrate the promising potential of multimodal depression detection. As research continues, integrating additional modalities like facial expressions or physiological data might lead to even more accurate and holistic depression screening tools.

3.6.2 Challenges and Considerations in Depression Detection

While the potential of using machine learning for depression detection is significant, there are crucial ethical and practical challenges to address:

3.6.2.1 Data Privacy Concerns

- **Data Collection and User Consent:** Collecting voice and text data, especially for individuals with depression, raises privacy concerns. Clear and informed user consent must be obtained, ensuring users understand how their data is collected, stored, and used. Transparency regarding data anonymization and security practices is essential [28].
- **Data Security and Potential Misuse:** The collected information can be highly sensitive, and vigorous security measures are required to ensure it from unauthorized get to or breaches [29]. There's also a risk of data misuse, and clear regulations are necessary to ensure the data is used solely for the aiming purpose of depression detection and not for commercial gain or discriminatory practices.

- **Potential for Social Stigma:** Depression is still a stigmatized condition, and individuals might hesitate to use tools that require sharing voice or text data for fear of discrimination [30]. Mitigating this stigma and promoting trust in the technology is crucial.

3.6.2.2 Model Generalizability Issues

The adequacy of machine learning models for depression detection can be hampered by issues related to generalizability:

- **Language Variations:** Models trained on data in one language might not perform well on data in another language due to variations in vocabulary, grammar, and emotional expression. Addressing this requires developing language-specific [31] models or models adaptable to diverse languages.
- **Cultural Contexts:** Cultural contexts heavily influence communication styles and expressions of emotions. A model trained on data from one culture might misinterpret vocal or textual cues from another culture, leading to inaccurate depression detection. Culturally sensitive model development and data collection are essential.
- **Communication Styles:** Individual communication styles vary greatly. Someone who is naturally quiet or speaks slowly might be misclassified as depressed [32]. Models need to be robust enough to account for these natural variations and avoid biased predictions.

These challenges highlight the importance of responsible development and deployment of machine learning tools for depression detection. Prioritizing user privacy, ensuring data security, and developing models that are generalizable across languages and cultures are critical steps towards creating trustworthy and effective solutions.

3.6.2.3 Role as Screening Tools

It's crucial to emphasize that machine learning and NLP-based tools for depression detection are envisioned as screening tools, not replacements for a formal clinical diagnosis [33]. Here's why:

- **Limited Scope:** These tools analyze text or voice information and recognize designs that could be indicative of discouragement. In any case, they cannot capture the full picture of an individual's mental health, as highlighted in [22]. A clinical diagnosis involves a comprehensive evaluation by a trained professional, considering factors like personal history, medical conditions, and symptom severity according to standardized criteria in resources like the DSM-5 [24].
- **Accuracy Considerations:** While machine learning models are constantly evolving, they are not perfect. There is always a risk of misclassification, including both false positives (identifying someone as depressed who isn't) and false negatives (missing someone who is actually depressed) [34]. A clinical diagnosis provides a more definitive assessment based on a broader range of information and reduces the risk of misclassification.
- **Importance of Human Interaction:** Depression is a complex condition, and individuals need personalized support and treatment plans. A clinical diagnosis lays the foundation for this by enabling a healthcare professional to establish a therapeutic relationship and develop a tailored approach for managing the depression, as emphasized in [35].

However, machine learning and NLP tools hold immense potential for:

- **Early Detection:** These tools can be used to screen large populations and identify individuals who might be at risk for depression, aligning with the goals of digital health tools for passive monitoring as discussed in [37]. Early detection is crucial for timely intervention and improved treatment outcomes.
- **Increased Access to Mental Health Resources:** By offering accessible screening tools, we can encourage individuals to seek help and potentially bridge the gap for those who might not readily access traditional mental health services [38]. This can address disparities in access to mental healthcare.
- **Monitoring Treatment Progress:** These tools can be used to screen changes in communication patterns over time, potentially aiding in tracking treatment advance and identifying areas where adjustments might be needed [40]. However, it is important to note that such applications are still beneath development, and assist investigate is required to approve their adequacy in this domain [39].

Machine learning and NLP offer valuable tools for depression screening. However, they should be seen as complementary to a comprehensive clinical evaluation and diagnosis by a healthcare professional. By leveraging these tools responsibly and ethically, we can create a more robust system for early detection, promoting timely access to mental health support and improved well-being for individuals experiencing depression.

3.6.3 Future Directions in Machine Learning for Depression Detection

Despite significant progress in developing machine learning (ML) models for depression detection using audio and text data, several critical research gaps persist, limiting the scalability and effectiveness of these systems in real-world applications. Some of the key areas that hold immense promise for the future are:

3.6.3.1 The Imperative of Explainable AI (XAI)

While traditional machine learning models excel at pattern recognition and prediction, they often operate as black boxes. In the realm of mental health, understanding how a model arrives at a depression prediction is critical for building trust and ensuring responsible use. Explainable AI (XAI) approaches empower us to:

- **Demystify Model Decisions:** By peering into the inner workings of the model, XAI techniques can reveal which features in the data (e.g., specific word choices in text analysis or prosodic patterns in voice analysis) have the strongest influence on the model's prediction of depression. This level of transparency allows researchers and mental health professionals to gain valuable insights into the factors the model deems most indicative of depression.
- **Safeguard Against Bias:** XAI can be a powerful tool for identifying and mitigating potential biases in the training data or the model itself. Biases can lead to misclassification for certain demographics or communication styles. By leveraging XAI techniques, researchers can identify these biases and implement strategies to reduce their impact, ensuring fairer and more generalizable models.
- **Boost User Confidence:** If users can understand the reasoning behind a depression prediction, it can foster trust and acceptance of the technology, especially in cases where further clinical evaluation is recommended. Transparency regarding the

model's decision-making process empowers users to engage more actively with the results and make informed decisions about their mental health.

3.6.3.2 Multimodal Detection for an integral View

As discussed earlier, the limitations inherent in analyzing text or voice data in isolation can be overcome by adopting multimodal approaches. Combining text analysis with voice analysis and even facial expression recognition offers a more comprehensive perspective on an individual's emotional state. Here's how exploring multimodality can enhance depression detection:

- **Bridging the Gaps:** Text analysis excels at capturing the content of communication, but it might miss subtle emotional cues conveyed through voice or facial expressions. Voice analysis, on the other hand, can reveal vocal characteristics associated with depression, but it lacks the context provided by textual content. By combining these modalities, we can create a more holistic picture, potentially leading to more accurate and nuanced depression detection.
- **Venturing Beyond Audio-Visual:** The realm of multimodality extends beyond audio and visual data. Research on incorporating physiological data like heart rate variability or skin conductance holds promise for further enriching the model's understanding of an individual's emotional state. By harnessing these additional modalities, we can potentially develop even more robust and comprehensive tools for depression detection.

3.6.3.3 Addressing Diversity and Generalizability

For machine learning models to have a truly global impact, ensuring generalizability across languages, cultures, and communication styles is paramount. Here's how we can move towards more inclusive and generalizable models:

- **Data Diversity is Key:** Developing models that are generalizable across diverse populations requires access to rich and comprehensive datasets that represent a wider range of languages, ethnicities, and communication styles. This might involve actively collecting data from underrepresented groups and employing techniques to reduce bias in model training.
- **Cultural Nuances Matter:** Cultural contexts heavily influence communication styles and expressions of emotion. A model trained primarily on data from one

culture might misinterpret vocal or textual cues from another culture, leading to inaccurate depression detection. Addressing this challenge necessitates incorporating cultural sensitivity into model development and data collection strategies.

- **Beyond Language Barriers:** Language variations can pose a significant challenge for generalizability. Developing language-agnostic models or models that can adapt to diverse languages is crucial for widening the reach of these tools.

By prioritizing these future directions, machine learning and NLP tools can evolve into even more powerful allies in the fight against depression. By fostering transparency, comprehensiveness, and inclusivity, these tools can empower early detection, promote timely access to mental health resources, and ultimately contribute to improved well-being for individuals experiencing depression on a global scale.

In the domain of mental health, machine learning (ML) and Natural Language Processing (NLP) have emerged as promising tools for depression detection using text and voice data. NLP techniques can analyze written communication to identify sentiment, word choice, and linguistic markers associated with depression. Voice analysis, on the other hand, can reveal vocal characteristics like pitch variations, speech rate, and prosody that might be indicative of the condition. Furthermore, combining these modalities (text and voice analysis) holds significant potential for improved accuracy and a more comprehensive understanding of an individual's emotional state. These advancements have the potential to revolutionize depression screening by facilitating early detection and encouraging help-seeking behavior. However, it is crucial to acknowledge and address limitations such as privacy concerns regarding data collection and security, as well as ensuring model generalizability across diverse languages, cultures, and communication styles. By prioritizing the development of explainable AI (XAI) for transparent decision-making and exploring the integration of additional modalities like facial expressions or physiological data, ML and NLP tools can evolve into even more robust and reliable allies in the fight against depression. Ultimately, these advancements have the potential to contribute significantly to a world where depression is effectively managed and treated.

CHAPTER FOUR

DEPRESSION DETECTION METHODOLOGY

4.1 Introduction

This thesis aims to recognize whether a person is depressed or not, and for that, a classification task is performed to predict who the depressed person is. The DAIC-WOZ dataset provides the binary classification of an individual's depression (depressed / not depressed) based on the severity of depression, measured using the Patients Health Questionnaire PHQ-8. The predictions are performed under the assumption that there are some similarities in the speech signals extracted from different people suffering from depression. For example, diminished, prosodic and monotonous speech is often strongly correlated with depression. the problem is a supervised binary classification, the following algorithms are to be implemented:



Figure 15: Selected models

The audio is preprocessed to obtain a more accurate representation of the participants' speech. Next, audio features are properly extracted using the Covarep toolbox. Finally, implemented in a list of machine learning models is implemented while the results of each models' predictions are compared and analyzed. Overall, the goal within this thesis is to compare Machine Learning and NLP method performances across different models' experimentation.

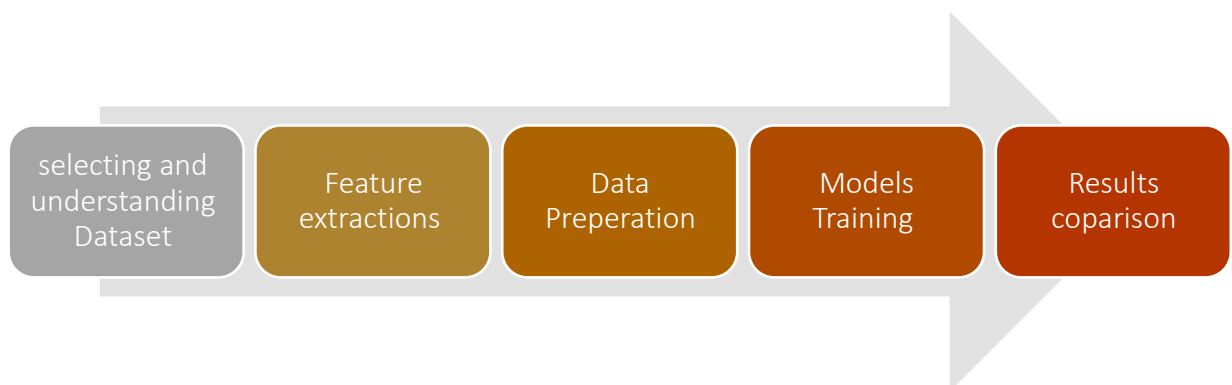


Figure 16: Study workflow

4.2 Dataset source

The Distress Analysis Interview Corpus (DAIC) is a comprehensive collection of clinical interviews aimed at diagnosing various psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD). These interviews were conducted as part of a broader initiative to develop a computerized interviewer capable of identifying both verbal and nonverbal indicators of mental illness.

Within the DAIC corpus, there exists a subset known as the DAIC-WOZ depression dataset [46]. This dataset comprises interviews conducted using a Wizard-of-Oz setup, where an animated virtual interviewer named Ellie interacted with participants under the control of a human interviewer situated in another room. The interviews were structured to elicit responses and behaviors indicative of mental health conditions, and data were collected in both audio and video formats.

In addition to audio and video recordings, participants also provided extensive questionnaire responses, offering further insights into their mental health status, symptoms, and demographic information. The collected data, including transcripts and annotations, encompass a wide range of verbal and nonverbal features associated with depression, anxiety, PTSD, and other mental health conditions [47].

Participants for these interviews were recruited through online ads posted on Craigslist.org, and the interviews themselves were conducted at the USC Institute for Creative Technologies (ICT) in Los Angeles, California. Fluent English speakers made up the participant pool, and the interviews, ranging from about 9 to 20 minutes, were conducted in English.

Before the interviews began, each participant completed the Patient Health Questionnaire (PHQ) on a computer. The interview process commenced with simple questions aimed at establishing a comfortable atmosphere. As the interview progressed, the questions became more specific, focusing on symptoms and events related to depression and post-traumatic stress disorder (PTSD). Finally, the interviews concluded with simpler questions once again, ensuring that participants did not leave in a distressed state of mind.

Participants were recorded using various devices, including a camera, high-quality close-talking microphone, and Kinect. Audio was captured through a head-mounted microphone (Sennheiser HSP 4-EW-3) at a sampling rate of 16kHz.

In the transcript file, sentences spoken by both the participant and the virtual assistant Ellie are documented. Each row of the transcript includes the start time (when the speaker begins speaking), stop time (when the speaker finishes speaking), speaker identification (Ellie or the participant), and the spoken sentence.

In total, the dataset includes audio recordings of 189 interviews, along with their corresponding transcripts and facial data. Additionally, each participant's PHQ-8 score, derived from their responses to the PHQ-8 questionnaire, is provided. The audio files are stored in raw format.

The package includes 189 folders of sessions 300-492. Certain sessions have been excluded for technical reasons. Data are grouped by session as below :

```
Pack\  
    300_P  
    301_P  
    ...  
    492_P  
    util  
    documents  
    train_split.csv  
    dev_split.csv  
    test_split.csv
```

Figure 17: Dataset content [45]

The dataset contains COVAREP, formant, and transcript feature files extracted from each interview using the COVAREP framework. These files encompass a variety of features obtained from both the participant's and the virtual interviewer's voices.

Furthermore, the dataset includes split files for training, development, and testing. The training and development splits present participant ID, PHQ-8 binary label, PHQ-8 score, participant gender, and responses to every question of the PHQ-8 questionnaire.

Conversely, the test split file only includes participant IDs and their genders. These split files aid in organizing the dataset for training, validation, and testing phases, ensuring reliable assessment and applicability of machine learning models.

Table 5: Train Development Test Set Attributes

	Train	Development	Test
Participant_ID	✓	✓	✓
Gender	✓	✓	✓
PHQ8_Binary labels	✓	✓	
PHQ8_Score	✓	✓	
PHQ8_NoInterest	✓	✓	
PHQ8_Depressed	✓	✓	
PHQ8_Sleep	✓	✓	
PHQ8_Tired	✓	✓	
PHQ8_Appetite	✓	✓	
PHQ8_Failure	✓	✓	
PHQ8_Concentrating	✓	✓	
PHQ8_Moving	✓	✓	

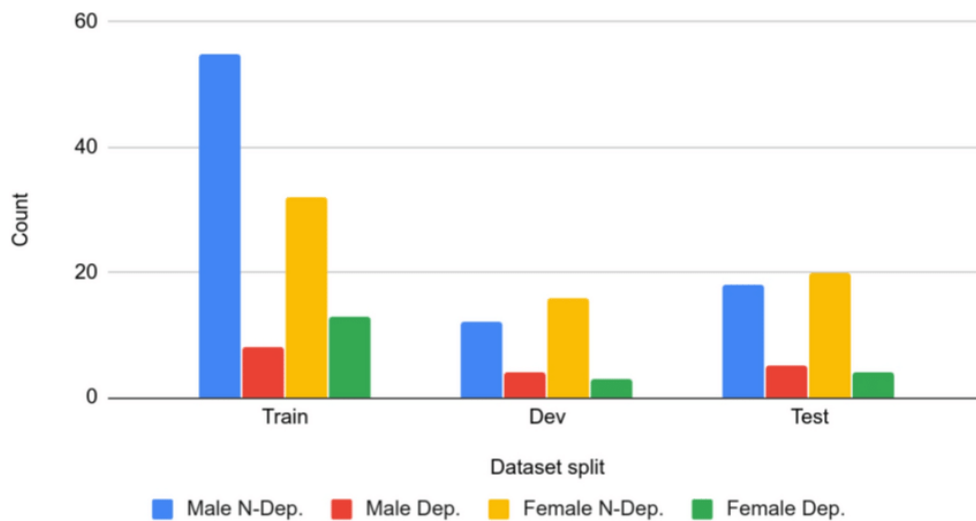


Figure 18: Data distribution of DAIC-WOZ depression database [48]

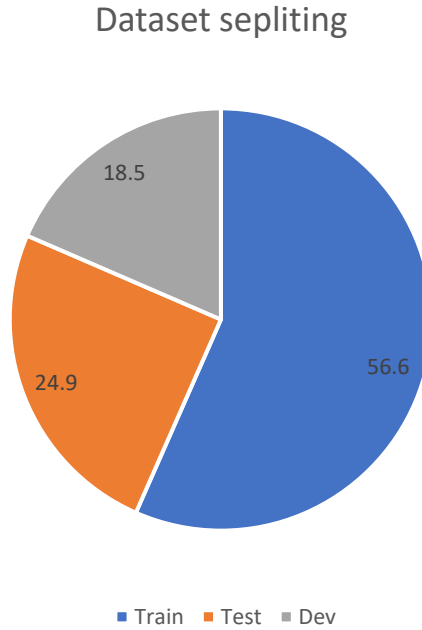


Figure 19: Dataset splitting

4.3 Feature Extraction

Audio features have already been extracted and are included in the DAIC-Woz dataset. These features were extracted using the COVAREP toolbox for speech analysis, employing well-validated and tested methods aimed at capturing various aspects of the speaker's voice.

Specifically, the extracted prosodic features encompass the Fundamental Frequency (F0) and Voicing Boundaries (VUV). Features related to voice quality include the Normalized Amplitude Quotient (NAQ), quasi-open quotient (QOQ), difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peakslope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd).

Moreover, spectral features comprise 25 Mel cepstral coefficients (MCEP0-24), Harmonic Model and Phase Distortion Mean (HMPDM0-24), and Harmonic Model and Phase Distortion Deviations (HMPDD0-12). These features collectively provide comprehensive insights into the prosodic characteristics, voice quality, and spectral attributes of the audio data.

4.3.1 Data Preprocess

The effectiveness of machine learning models heavily depends on the quality of the data used. For the DAIC-WOZ dataset, a meticulous process was followed to ensure the data collected was clear and free from significant noise or interference. Although great care was taken, perfect data is rarely achievable. Given that the interviews were conducted by a virtual interviewer, occasional technical issues were anticipated. The dataset documentation outlines instances where interruptions occurred—seven interviews in total—which were subsequently discarded.

The audio recordings originally captured the voices of both the virtual interviewer, Ellie, and the participants. Since Ellie’s contributions are unrelated to measuring the severity of participants’ depression, including her speech would only introduce unnecessary interference. To avoid this, the dataset offers transcripts in CSV format, clearly identifying each speaker’s segments and their corresponding timestamps, as illustrated in Figure 20. This allows for easy isolation of the participants' speech, ensuring that the final audio data remains relevant and free from extraneous input.

start_time	stop_time	speaker	value
36.588	39.668	Ellie	hi i'm ellie thanks for coming in today
39.888	43.378	Ellie	i was created to talk to people in a safe and secure environment
43.728	48.498	Ellie	think of me as a friend i don't judge i can't i'm a computer
49.188	52.388	Ellie	i'm here to learn about people and would love to learn about you
52.658	58.958	Ellie	i'll ask a few questions to get us started and please feel free to tell me anything your answers are totally confidential
60.028	61.378	Ellie	how are you doing today
62.328	63.178	Participant	good
63.798	64.738	Ellie	that's good
65.858	67.528	Ellie	where are you from originally

Figure 20: Transcript sample [46]

4.4 Classifiers

4.4.1 CNN

Convolutional Neural Networks (CNNs) are a class of deep learning algorithms specifically designed for processing structured grid-like data, such as images. They have demonstrated remarkable performance in various computer vision tasks, including image classification, object detection, and image segmentation.

The fundamental building blocks of CNNs include convolutional layers, pooling layers, and fully connected layers. Let's delve into each of these components along with the mathematical equations that govern their operations:

1. Convolutional Layers:

The primary operation in this layer is the convolution operation, which involves sliding a filter (or kernel) over the input image and computing the dot product of the filter with local regions of the input image can be represented as follows:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n) \quad (26)$$

2. Activation Function

After the convolution, an activation function is applied to introduce non-linear properties into the network. The most common activation function in CNNs is the Rectified Linear Unit (ReLU), defined as:

$$f(x) = \max(0, x) \quad (27)$$

3. Pooling Layers:

Pooling layers down sample the feature maps obtained from the convolutional layers to reduce computational complexity and extract the most relevant information. Max pooling is a common pooling operation, which selects the maximum value within each pooling window. The mathematical expression for max pooling can be represented as:

$$P(i, j) = \max_{m, n} (S(m, n)) \quad (28)$$

4. Fully Connected Layers:

Fully connected layers perform classification based on the features extracted by the convolutional and pooling layers. The output of the fully connected layer is computed using the following equation:

$$y = W_x + b \quad (29)$$

Training at CNN involves optimizing the parameters (weights and biases) of the network to minimize a loss function, often defined as the cross-entropy loss for

classification tasks. This optimization process is typically performed using gradient descent-based algorithms such as stochastic gradient descent (SGD) or its variants.

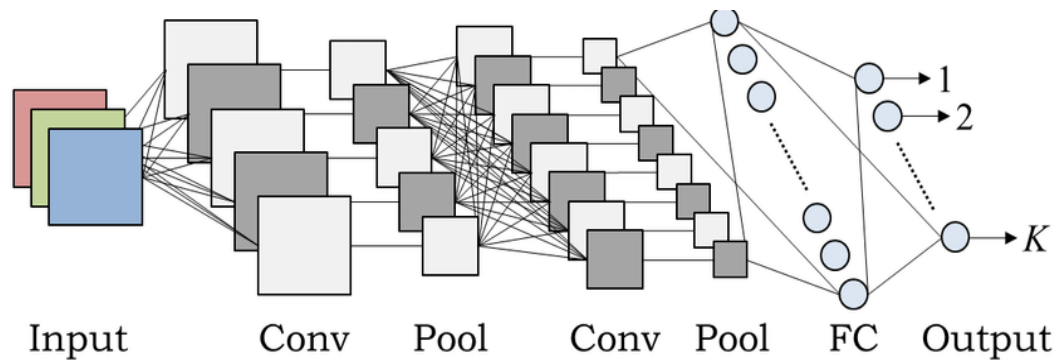


Figure 20: CNN Architecture [54]

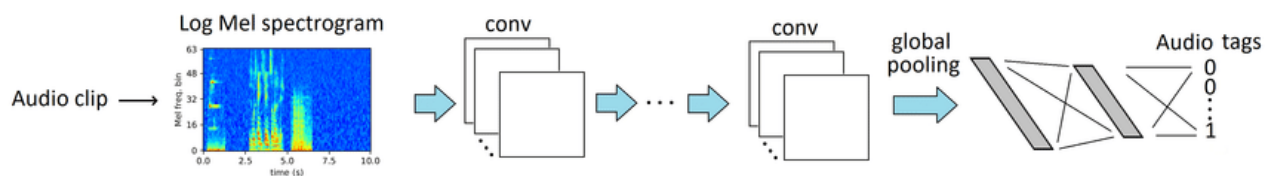


Figure 21: Decoding Audio in the CNN Architecture [55]

4.4.2 RF

Random Forest is a versatile and powerful ensemble learning algorithm that is widely used for both classification and regression tasks. It works by constructing a multitude of decision trees during training and outputting the mode (classification) or mean prediction (regression) of the individual trees.

Random forest classifiers are often considered to be among the most accurate and robust machine learning algorithms, and they are widely used in a variety of applications, including image and text classification, fraud detection, and predictive maintenance.

A random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement.

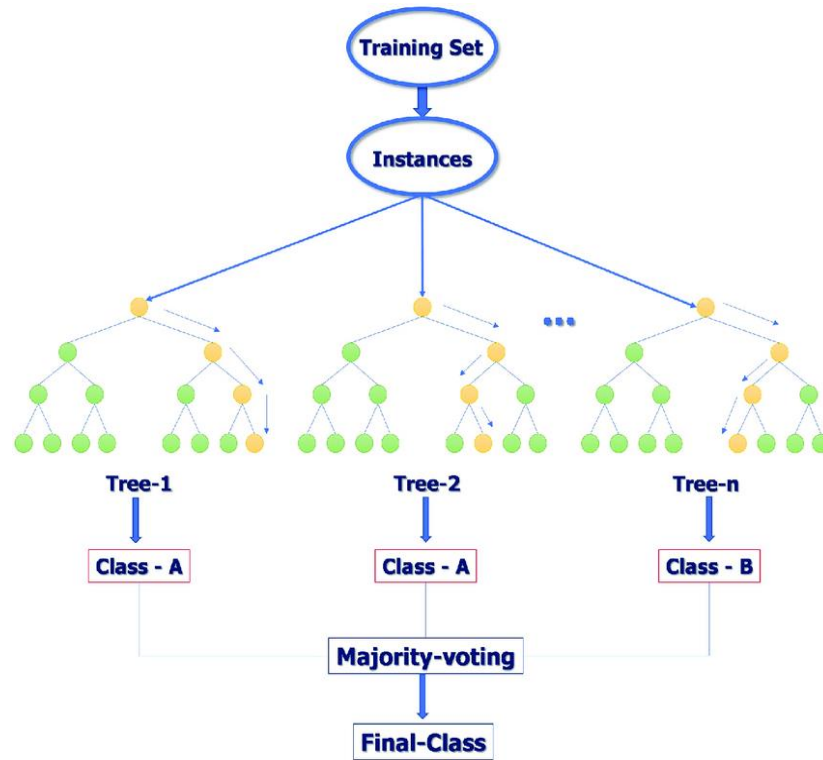


Figure 22: Random Forest Algorithm [56]

4.4.3 SVM

A Support Vector Machine (SVM) is a supervised machine learning algorithm developed primarily for classification tasks. It operates by establishing an optimal hyperplane which maximizes the separation margin between different classes in a multidimensional space. This concept was first introduced by Vladimir N. Vapnik and his colleagues in their 1995 paper, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing".

In practice, SVMs classify data by identifying the hyperplane that maximizes the distance between the nearest data points of opposing classes, which are referred to as support vectors. These support vectors define the boundaries of the margin and are critical in determining the best decision boundary that separates the classes. Depending on the number of features, the hyperplane can appear as a line in two-dimensional space or a higher-dimensional plane.

SVMs are adept at handling both linear and nonlinear classification. For non-linear classification, the algorithm employs kernel functions to project the input data into a higher-dimensional space where a linear separator is feasible. This method, known as the "kernel trick," allows the SVM to effectively deal with more complex datasets. The choice of the kernel function—linear, polynomial, radial basis function (RBF), or sigmoid—depends on the dataset's characteristics and the specific requirements of the task. This versatility and robustness in maximizing margins make SVMs highly effective for a wide range of classification challenges, allowing them to generalize well to new, unseen data.

The key aspects of the SVM algorithm:

- Linear SVMs are best suited for datasets that are linearly separable, meaning the data can be divided into different classes without needing transformation. The decision boundary and support vectors visually resemble a wide street, an analogy used by Professor Patrick Winston from MIT to describe this quadratic optimization problem as "fitting the widest possible street." Mathematically, the separating hyperplane in a Linear SVM is defined by the equation:

$$k(x_i, x_j) = x_i * x_j \quad (30)$$

- Support Vectors: In SVM, only a subset of the training data points, called support vectors, influence the position and orientation of the hyperplane. These are the data points that are closest to the decision boundary or lie within the margin. The rest of the data points do not affect the hyperplane's position.
- Kernel Trick: SVM can efficiently handle non-linear classification tasks by mapping the input features into a higher-dimensional space using kernel functions. This allows SVM to find a linear decision boundary in the transformed space, even when the original feature space is not linearly separable. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels.
- Margin Maximization: SVM aims to maximize the margin, which represents the distance between the decision boundary and the nearest data points of each class. By maximizing the margin, SVM improves its generalization performance and reduces the risk of overfitting.

- **Regularization Parameter:** SVM includes a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing the classification error on the training data. A smaller value of C results in a larger margin but may lead to misclassification of some training examples, while a larger value of C allows for more flexibility in the decision boundary but may increase the risk of overfitting.
- **Multi-Class Classification:** SVM can be extended to handle multi-class classification tasks using techniques such as one-vs-one (OvO) or one-vs-rest (OvR) strategies, where multiple binary classifiers are trained to distinguish between pairs of classes.

The formula for SVMs can be mathematically described as:

$$\min \frac{1}{2} w^T w + c \sum_{i=1}^n \zeta_i \quad (31)$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i \quad (32)$$

4.4.4 LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to address the limitations of traditional RNNs in capturing long-term dependencies in sequential data. Unlike standard RNNs, which suffer from the vanishing gradient problem and struggle to retain information over long sequences, LSTMs are capable of learning and remembering information over extended time periods.

The key innovation of LSTMs lies in their memory cells, which maintain a cell state that can be updated or modified selectively using gating mechanisms. These gating mechanisms, including input gates, forget gates, and output gates, control the flow of information into, out of, and within the memory cell, enabling LSTMs to learn and retain relevant information while discarding irrelevant information.

The LSTM architecture consists of several components:

- 1- **Cell State:** The cell state serves as the memory of the LSTM and can theoretically maintain information over long sequences.
- 2- **Gates:** LSTMs use three types of gates to regulate the flow of information: input gates, forget gates, and output gates. These gates, which are sigmoid neural networks, decide which information should be let through and which should be discarded.
- 3- **Hidden State:** The hidden state, or output, of the LSTM is computed based on the cell state and input at each time step. It represents the LSTM's learned representation of the input sequence.

During training, LSTMs learn to update their parameters (weights and biases) using backpropagation through time (BPTT) or other optimization algorithms. By selectively updating the cell state and hidden state based on the input and previous states, LSTMs are able to capture long-term dependencies in sequential data and perform effectively on tasks such as language modeling, speech recognition, machine translation, and time series prediction.

4.4.5 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking natural language processing (NLP) model developed by Google AI Language in 2018. It represents a significant advancement in pretraining techniques for NLP tasks by introducing bidirectional context understanding.

Some of the key aspects of the BERT model are:

- 1- **Bidirectional Context Understanding:** Unlike previous NLP models that relied on unidirectional language models (LMs), which process text sequentially in one direction, BERT introduced a bidirectional approach. It is pre-trained on large text corpora using masked language model (MLM) and next sentence prediction (NSP) objectives, enabling it to capture context from both left and right contexts simultaneously.

- 2- Transformer Architecture: BERT is built upon the transformer architecture, which has proven to be highly effective for sequence-to-sequence tasks. The transformer architecture consists of self-attention mechanisms and feed-forward neural networks. It allows BERT to capture long-range dependencies and relationships between words in a sentence.

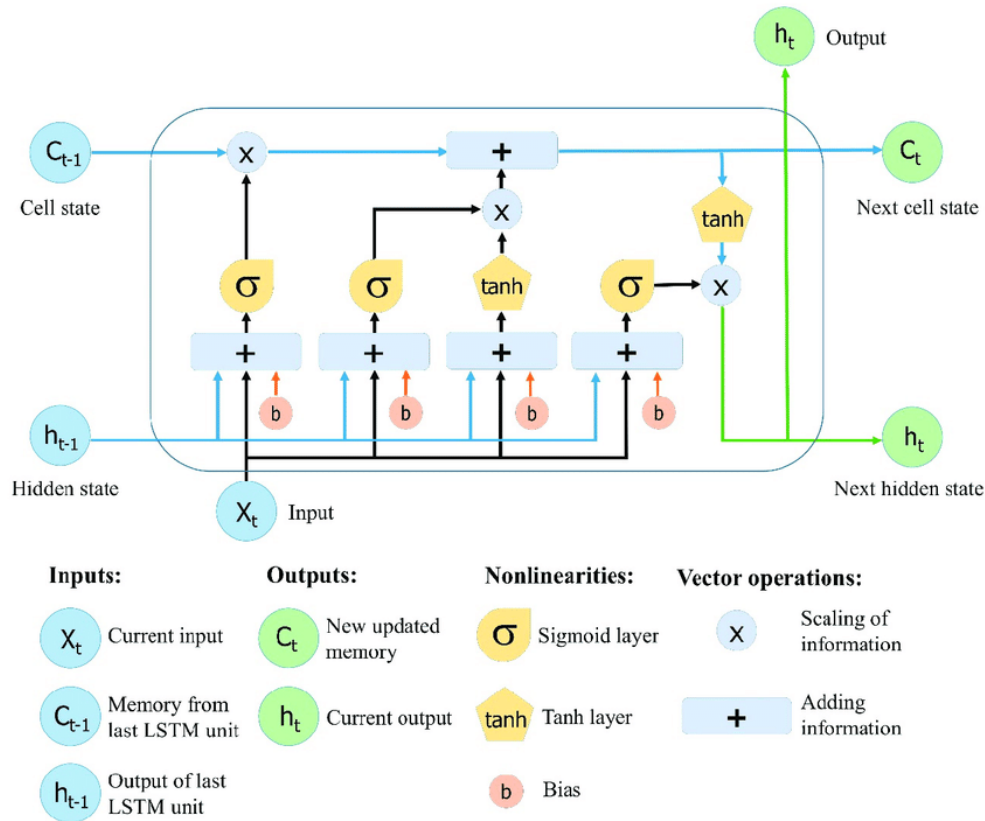


Figure 23: The Long Short-Term Memory (LSTM) Architecture [58]

In training language models, a key challenge lies in defining an appropriate prediction objective. Many models are designed to predict the next word in a sequence which adopts a unidirectional approach that inherently limits the ability to fully grasp context. To address this issue, BERT employs two distinct training strategies:

1. Masked Language Model (MLM):

In the MLM approach, 15% of the words in each input sequence are replaced with a [MASK] token before the sequence is fed into BERT. The model's task is to predict the original words that were masked, relying on the contextual information provided by the surrounding unmasked words. Technically, this prediction involves:

- Adding a classification layer on top of the encoder's output.
- Multiplying the output vectors by the embedding matrix, which transforms them into the dimensions of the vocabulary.
- Using softmax to calculate the probability distribution over the entire vocabulary for each word.

2. Next Sentence Prediction (NSP) :

The NSP strategy involves training BERT on pairs of sentences to determine whether the second sentence logically follows the first in the original text. In the training data, 50% of the pairs are in the correct sequence, while the remaining 50% are randomly paired, with the second sentence likely unrelated to the first. The input preparation for NSP includes:

- Inserting a [CLS] token at the beginning of the first sentence and a [SEP] token at the end of each sentence to signal the boundary between them.
- Assigning each token a sentence embedding, marking it as part of either Sentence A or Sentence B, with a vocabulary limited to two labels.
- Applying a positional embedding to each token to indicate its position in the sequence, based on the principles outlined in the Transformer model.

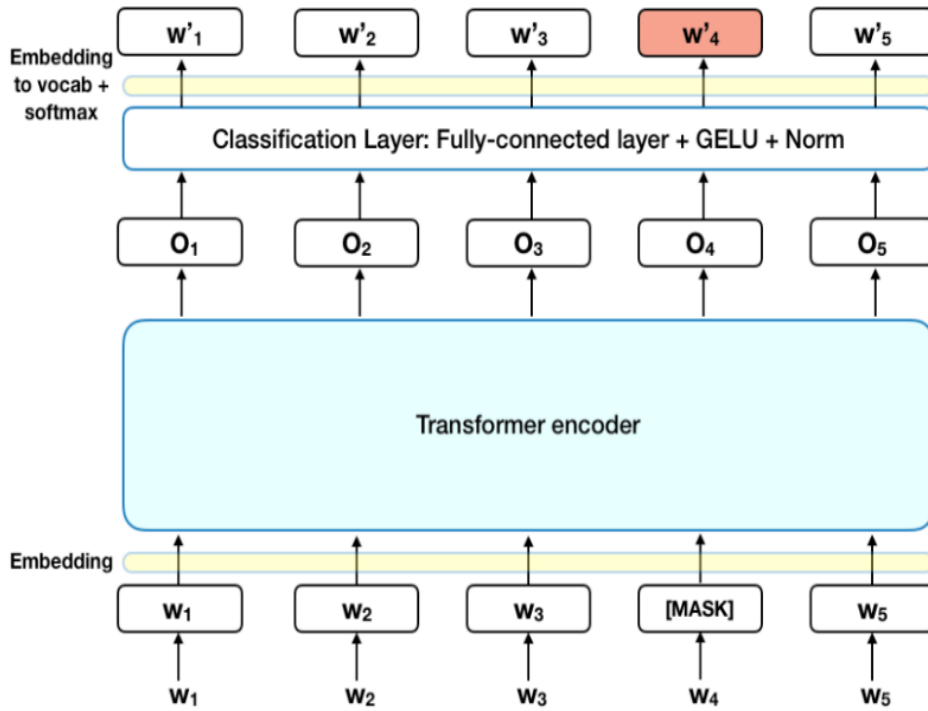


Figure 24: MLM structure of the BERT model [59]

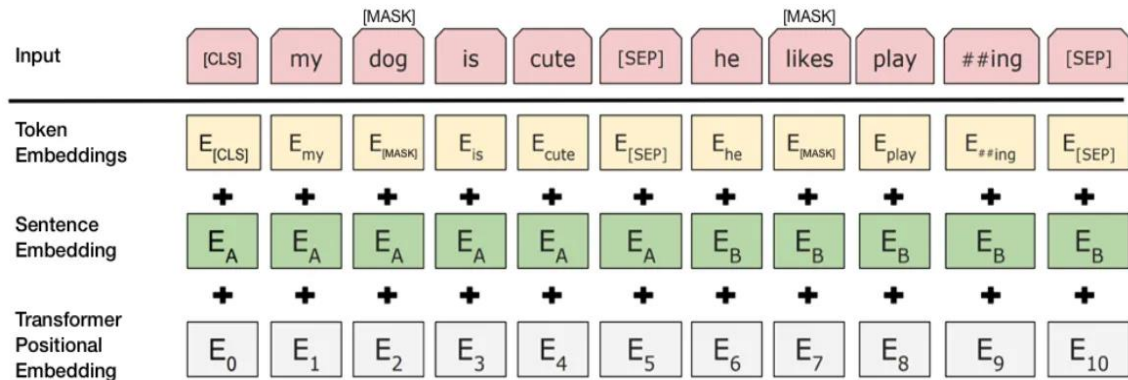


Figure 25: BERT model [66]

To evaluate whether the second sentence follows the first, the BERT model processes the entire input sequence through its Transformer architecture. The output corresponding to the [CLS] token is then passed through a classification layer, which converts it into a 2x1 vector using learned weights and biases. The likelihood of the sequence being the next sentence is calculated using a softmax function.

During the training of the BERT model, both MLM and NSP tasks are simultaneously trained. The objective is to minimize the combined loss from these two strategies,

optimizing the model's performance on understanding and predicting sentence relationships.

BERT Fine-Tuning: After pretraining on large corpora, BERT can be fine-tuned on specific downstream tasks with task-specific labeled data. By fine-tuning, BERT adapts its representations to the task at hand, achieving state-of-the-art performance on various NLP tasks such as sentiment analysis, named entity recognition, and machine translation.

Model Variants: BERT comes in different sizes, including BERT-base, BERT-large, and BERT-xlarge. The larger variants have more parameters and are capable of capturing more complex patterns in the data but require more computational resources.

In BERT Model Model size significantly impacts performance, even at a large scale. For example, BERT_large, which comprises 345 million parameters, is the most extensive model in its category and consistently outperforms BERT_base on smaller tasks. Despite sharing the same architecture, BERT_base has "only" 110 million parameters.

With sufficient training data, increasing the number of training steps correlates with higher accuracy. For instance, in the MNLI task, BERT_base shows a 1.0% improvement in accuracy when trained for 1 million steps (with a batch size of 128,000 words) compared to 500,000 steps at the same batch size.

Although BERT's bidirectional training method (MLM) converges more slowly than traditional left-to-right approaches—due to only 15% of words being predicted in each batch—it still surpasses left-to-right training in performance after a relatively small number of pre-training steps.

BERT has had a profound impact on the field of NLP, significantly advancing the state of the art in various language understanding tasks. Its ability to generate high-quality contextualized word embeddings has made it a go-to choice for many NLP applications and has paved the way for further research in transformer-based models.

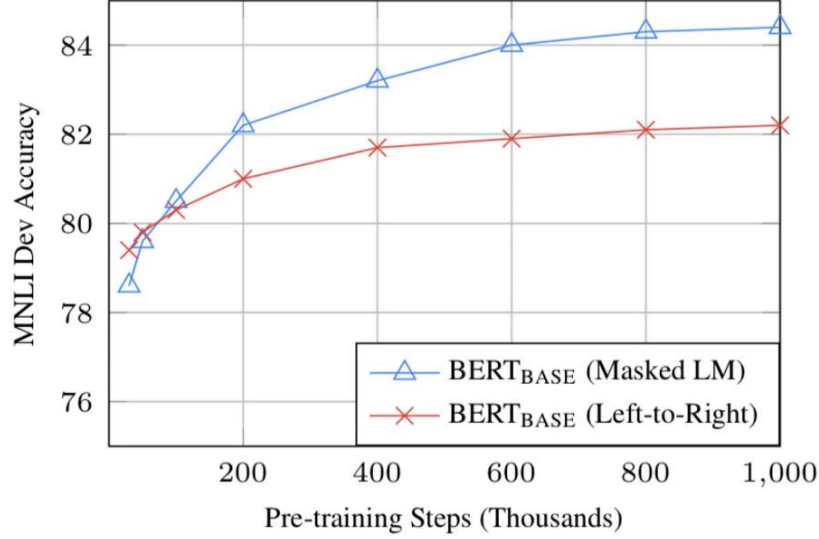


Figure 26: BERT training strategies [66].

4.5 Evaluation Measures

Evaluation measures in machine learning (ML) are critical for assessing the effectiveness of models. They help to quantify how well a model performs on specific tasks, such as classification, regression, or clustering. Different tasks require different evaluation metrics, and choosing the right metric is vital for interpreting model performance accurately and making informed decisions about which models to deploy.

Accuracy: The most straightforward metric measures the proportion of correctly predicted observations to the total observations. It's suitable for balanced datasets but can be misleading in the presence of imbalanced classes. It is the sum of True Positives (TP) and True Negatives (TN) divided by the sum of the total – True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Accuracy, as well, is insufficient when dealing with imbalanced classes.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (33)$$

Precision and Recall: Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall (Sensitivity) is the ratio of correctly predicted

positive observations to all observations in the actual class. These metrics are particularly useful for imbalanced datasets.

Precision: measures the extent of error introduced by False Positives (FP). It indicates how many predictions in the positive class were correct. However, this metric can be unreliable for evaluating the performance of models on imbalanced datasets.

$$Precision = \frac{TP}{TP+FP} \quad (34)$$

Recall : known as Sensitivity, contrasts with precision as it reflects the extent of error due to False Negatives (FNs). It measures the proportion of actual positive instances that were correctly classified as positive.

$$Recall = \frac{TP}{TP+FN} \quad (35)$$

F1 Score : The harmonic mean of precision and recall. It is a better measure than accuracy for imbalanced datasets because it takes both false positives and false negatives into account. achieving its best value at 1 which represents perfect precision and recall, and its worst at 0.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (36)$$

ROC-AUC Score: Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate at various threshold settings. The Area Under the Curve (AUC) represents a model's ability to discriminate between positive and negative classes.

Confusion Matrix : A confusion matrix is a table used to describe the performance of a classification model on a set of data for which the true values are known. It presents the counts of true positives, true negatives, false positives, and false negatives, which help to measure the accuracy of a classification. Confusion matrices are typically arranged:

- True Positive (TP): The count of instances that were correctly predicted as positive.
- False Positive (FP): The count of instances that were incorrectly predicted as positive.

- True Negative (TN): The count of instances that were correctly predicted as negative.
- False Negative (FN): The count of instances that were incorrectly predicted as negative

Table 6: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

In the predicted outcomes, values are categorized as "positive" or "negative" based on the model's predictions. "TP" represents true positives, indicating the count of positive instances correctly identified by the model. "TN" denotes true negatives, reflecting the count of negative instances correctly identified. "FP" stands for false positives, showing the count of negative instances mistakenly classified as positive, and "FN" refers to false negatives, representing the count of positive instances mistakenly classified as negative. In this context, "negative" corresponds to non-depressed participants, while "positive" pertains to those identified as depressed.

4.6 Data Preparation

Initially, data is imported from various CSV files, which include depression levels and identifiers for audio files stored in the form "_COVAREP.csv". These files contain vocal features extracted for analysis. The data from these files is subjected to preprocessing methods such as:

- **Zeroing:** Modifying specific rows based on conditions to aid in focusing on relevant features.
- **Scaling Down:** Reducing the size of the dataset to manage computational load.
- **Decreasing Size:** Ensuring uniformity in data length across samples to stabilize the input to models.

This helps for handling large datasets effectively and ensuring the models receive clean and manageable data inputs. This approach reflects a preparation to manage the size

and quality of data, making sure that only the most relevant temporal segments remain for model training.

After training and testing the models, the results are often quantified using several standard metrics:

- **Accuracy:** Measures the percentage of total correct predictions.
- **Precision:** Indicates the proportion of positive identifications that were actually correct.
- **Recall (Sensitivity):** Measures the proportion of actual positives that were correctly identified.
- **F1-Score:** Harmonic means of precision and recall, providing a balance between the two in cases of uneven class distributions.

```
X_temp = data.iloc[:, :].values
X_temp = makedata(X_temp) # Assuming makedata is defined elsewhere
X_temp = X_temp[X_temp.shape[0]-size:]
X_train.append(X_temp)
Y_train.append(train[i][1])
```

4.7 Audio Models

4.7.1 CNN

The core operation in our CNN involves convolution, applied as follows:

$$f(x) = (W * x + b) \quad (37)$$

Where W is the weight matrix, x the input, b the bias, and $*$ denotes the convolution operation.

We used 2 CNN models and the results of the second one overclass the results of first one. We achieved these results by making a shift in the architecture of the CNN model. The CNN designed for audio data analysis uses 1D convolution layers, which are ideal for extracting temporal features from sequential data such as audio.

1. **Convolution Layers:** Multiple layers with decreasing filter sizes extract various levels of features from raw audio data. Starting with broader features and moving to more specific features in deeper layers.
2. **Pooling Layers:** MaxPooling layers are used after convolution layers to reduce the dimensionality, which helps in reducing computation as well as in making feature detectors more invariant to slight changes.
3. **Flattening:** Before connecting to dense layers, the data is flattened from a matrix to a vector.
4. **Dense Layers:** Fully connected layers interpret the features extracted and flattened by previous layers to make final predictions. A dropout layer is included to reduce overfitting by randomly dropping units from the neural network during training.
5. **Output Layer:** Uses a sigmoid activation function to output a probability, indicating the likelihood of depression.

CNN is trained using batches of preprocessed audio data, labeled as either indicating depression or not. The training process involves adjusting weights through backpropagation, based on the loss calculated by the binary cross entropy function and

monitored using accuracy. The results show how well the model can generalize from training data to unseen data. Based on the output.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 39991, 60)	44460
max_pooling1d_3 (MaxPooling1D)	(None, 13330, 60)	0
conv1d_4 (Conv1D)	(None, 13326, 30)	9030
max_pooling1d_4 (MaxPooling1D)	(None, 4442, 30)	0
conv1d_5 (Conv1D)	(None, 4438, 15)	2265
max_pooling1d_5 (MaxPooling1D)	(None, 1479, 15)	0
flatten_1 (Flatten)	(None, 22185)	0
dropout_1 (Dropout)	(None, 22185)	0
dense_2 (Dense)	(None, 128)	2839808
dense_3 (Dense)	(None, 1)	129

Figure 27: CNN model

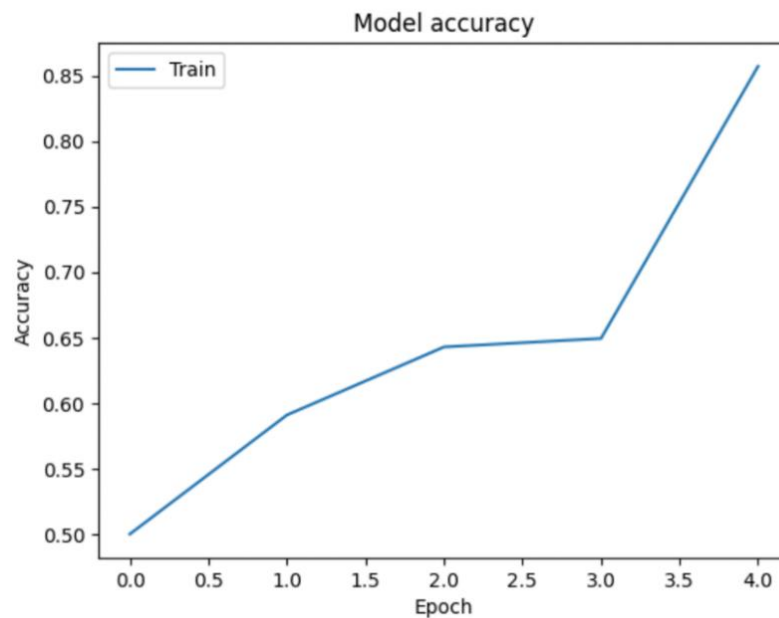


Figure 28: CNN accuracy plot

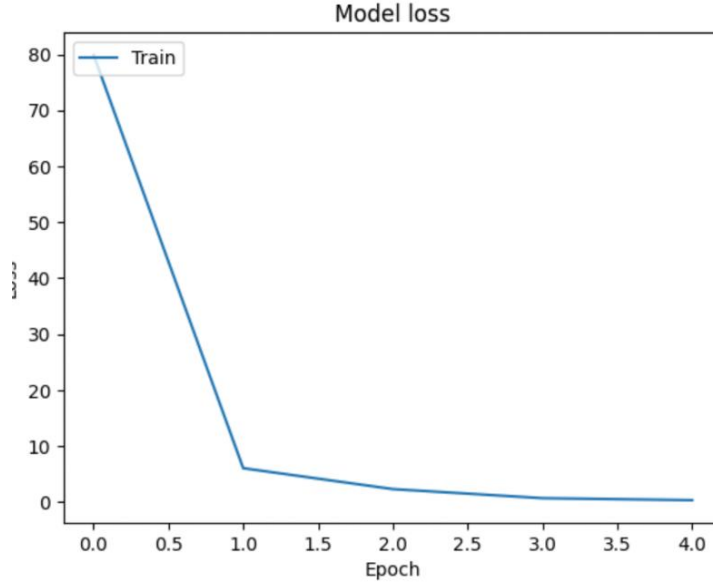


Figure 29: CNN loss plot

The CNN model architecture is carefully constructed to handle audio data effectively by extracting meaningful features through convolution and pooling layers and then using dense layers to make predictions. The use of dropout helps in avoiding overfitting, ensuring that the model generalizes well on unseen data.

```

5/5 [=====] - 4s 702ms/step
Y_pred: (154, 1)
      precision    recall  f1-score   support

     0.0         0.63      1.00      0.77         77
     1.0         1.00      0.40      0.57         77

 accuracy          0.70         154
 macro avg         0.81         0.70      0.67         154
 weighted avg         0.81         0.70      0.67         154

```

Figure 30: CNN model classification report

4.7.2 LSTM

LSTM units use the following update equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (38)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (39)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (40)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (41)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (42)$$

$$h_t = o_t * \tanh(C_t) \quad (43)$$

Where σ is the sigmoid function, C_t is the cell state, and h_t is the hidden state.

LSTM are designed to address the issue of long-term dependencies, where the gap between the relevant information and the point where it is needed becomes large. They are particularly useful for analyzing sequences like audio recordings, where understanding the context from long sequences is crucial for accurate predictions.

4.4.2.1 Model Configuration

1. **Input Layer:** The input layer receives sequences of features. For us the number of features extracted from the audio data (40000 timesteps with 74 features each).
2. **LSTM Layers:** Multiple LSTM layers can be stacked for better feature extraction from complex sequences. Each LSTM layer can return sequences to the next layer
3. **Dropout Layers:** Used intermittently between LSTM layers to prevent overfitting by randomly dropping a fraction of the units during the training process.
4. **Output Layer:** A dense layer with a sigmoid activation function outputs a probability indicating the likelihood of depression.

4.7.2.2 LSTM Comparison

We've implemented four different LSTM models that are included in the code file, each with unique architectures and configurations. They differ in their architecture and the results achieved.

Out of all the models we have built, the two LSTM models built on the COVAREP file ended up giving results. They are different significantly in their architecture and performance, which likely contribute to their respective results. Here's a detailed comparison based on their architectures and performance metrics:

4.7.2.3 Architectural Difference

I. Model 1:

- Units: 64 LSTM units.
- Layers: Includes one LSTM layer followed by a Dropout layer for regularization and a final Dense layer for output.
- Complexity: Lower, with fewer units and layers. This setup may struggle with capturing more complex patterns in data but requires less computational power.

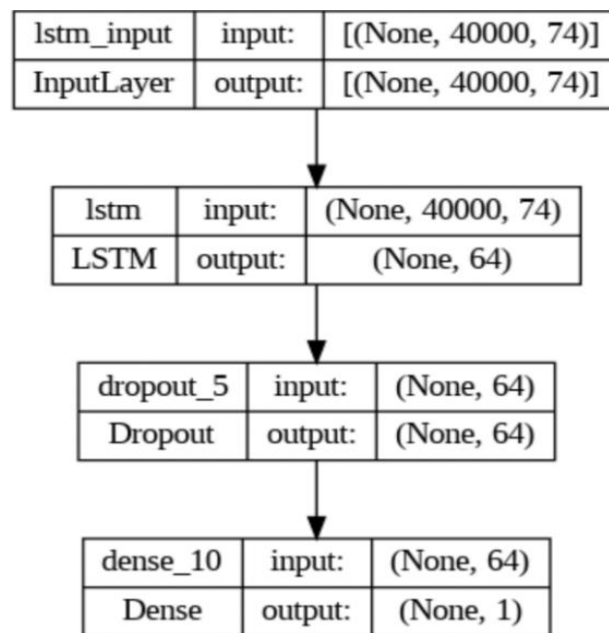


Figure 31: LSTM model 1 summary

II. Model 2:

- Units: 32 LSTM units, half as many as Model 1.
- Layers: Similar to Model 1 with one LSTM layer, one Dropout layer, and one Dense output layer.
- Complexity: Even lower than Model 1, further reducing the model's ability to learn complex patterns but increasing efficiency in terms of computation.

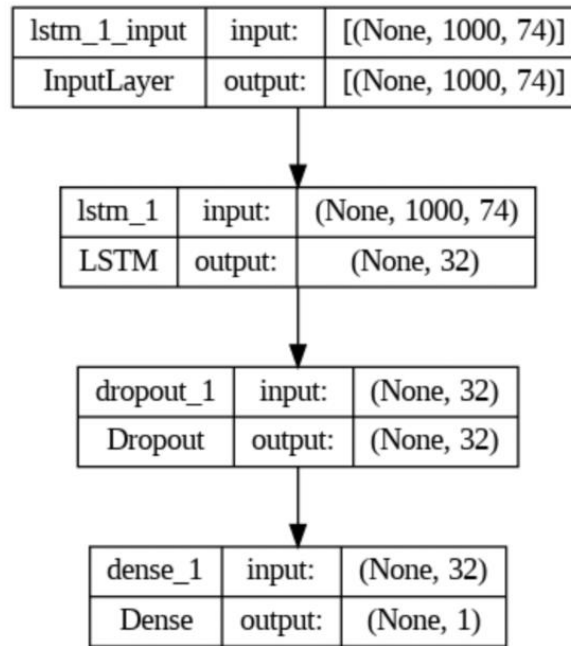


Figure 32: LSTM model 2 summary

The LSTM model is trained using the preprocessed and structured data of COVAREP CSV file, fitting the model to the labels indicating the presence or absence of depression. The training process involves 10 epochs, where the model learns by adjusting weights to minimize the loss function.

4.7.2.4 Performance Results

I. Model 1:

- **Validation Accuracy:** Reached a high of about 77.42% during training, but the test accuracy was only about 52.17%. This discrepancy suggests that the model might be overfitting to the training data or failing to generalize well to unseen data.
- **Classification Metrics:** Shows a significant bias towards predicting class 1 (depressed), with a high recall for class 1 but very low precision, indicating many false positives.

LSTM model offers a dynamic approach to understanding temporal relationships in audio data, crucial for detecting patterns indicative of depression.

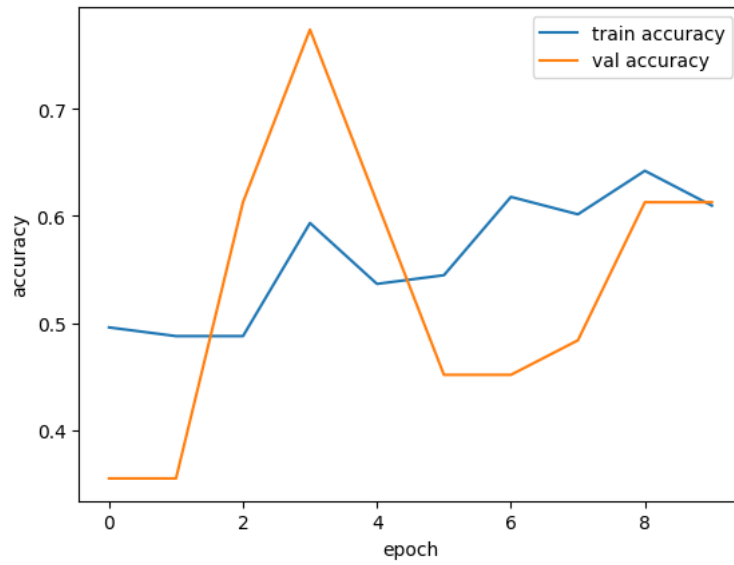


Figure 33: LSTM model 1 accuracy plot

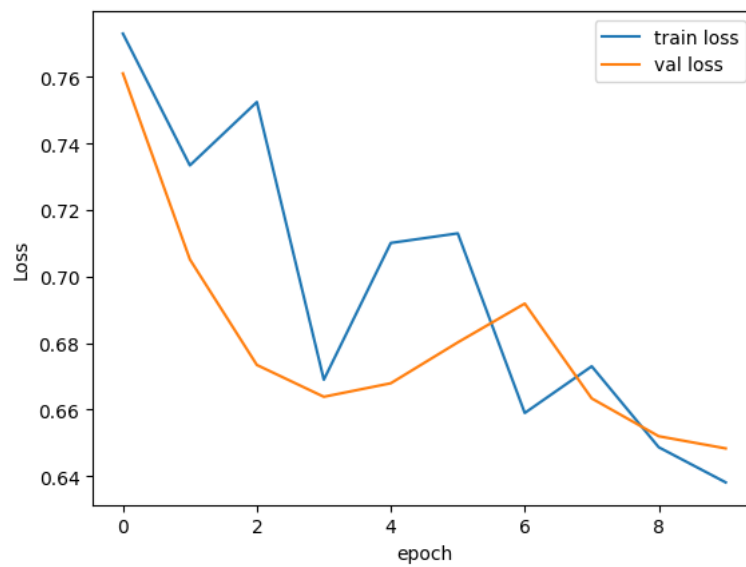


Figure 34: LSTM model 1 loss plot

Classification Report:				
	precision	recall	f1-score	support
0	0.81	0.41	0.54	32
1	0.37	0.79	0.50	14
accuracy			0.52	46
macro avg	0.59	0.60	0.52	46
weighted avg	0.68	0.52	0.53	46

Figure 35: LSTM model 1 classification report

II. Model 2:

- **Validation Accuracy:** Peaked at about 89.74% during training, indicating that it was effective at learning from the training data compared to Model 1. Classification Metrics: Showed a much better balance in precision and recall between the two classes, achieving a higher overall accuracy of 87% on the test set.

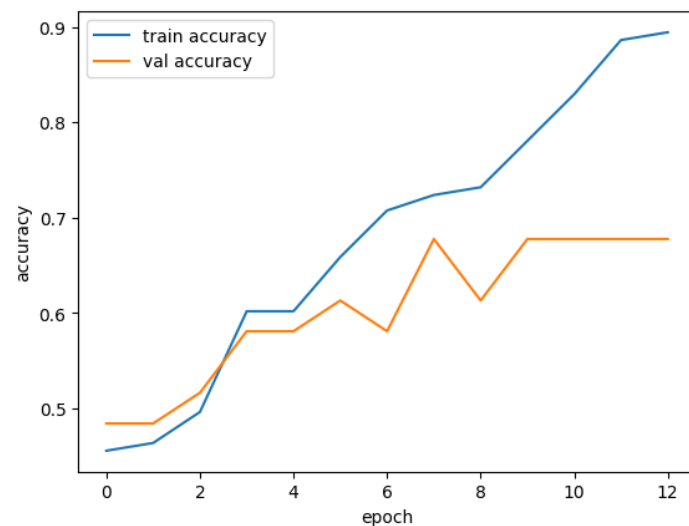


Figure 36: LSTM model 2 accuracy plot

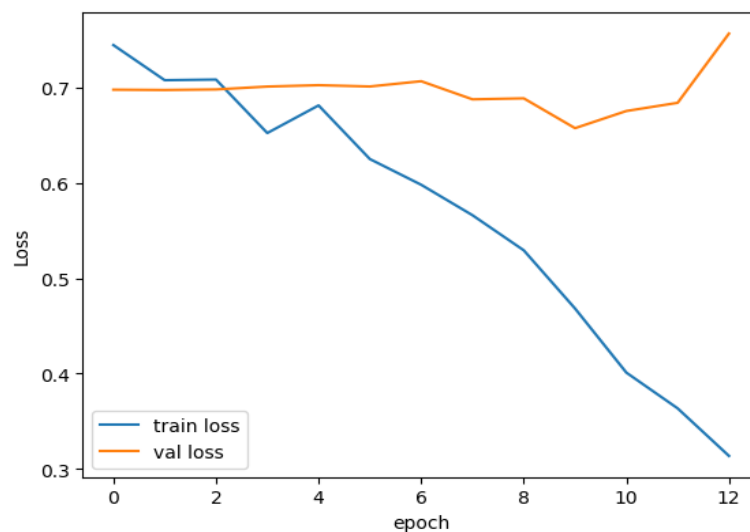


Figure 37: LSTM model 2 loss plot

	precision	recall	f1-score	support
Non-Depressed	0.94	0.78	0.85	74
Depressed	0.82	0.95	0.88	76
accuracy			0.87	150
macro avg	0.88	0.87	0.87	150
weighted avg	0.88	0.87	0.87	150

Figure 38: LSTM model 2 classification report

4.7.2.5 Analysis of Performance

The key differences in performance can be attributed to several factors:

- **Model Complexity vs. Data Complexity:** Model 1, with more units, might be slightly overfitting, especially given the small improvement over the epochs. It learned the training data well but failed to generalize this knowledge to the test data. Model 2, despite its simpler architecture, seems better at generalizing, possibly because it avoided overfitting simpler patterns present in the COVAREP feature set.
- **Training Efficiency:** Model 2, with fewer units, would have required less computational resources, making it more efficient. Its simpler architecture seems to be a better fit for the complexity of the data, avoiding overfitting and thus performing better on unseen test data.
- **Data Preprocessing and Feature Relevance:** Both models used COVAREP files, which include various acoustic features extracted from speech. The performance suggests that simpler models may be more suitable if the feature set does not contain highly complex patterns or if the most relevant features do not require complex models to be interpreted effectively.

Model 2's superior test performance despite its simpler architecture suggests it is a more suitable choice for this particular dataset and task. It managed to achieve higher accuracy with fewer resources and avoided the overfitting apparent in Model 1. This highlights the importance of matching model complexity to data complexity and the potential benefits of simpler models when dealing with datasets where the signal-to-noise ratio does not warrant a more complex approach.

4.7.3 RF Classifier

The Random Forest classifier is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes of the individual trees. Random Forests are known for their high accuracy, robustness, and ease of use. They automatically provide estimates of feature importance and can handle unbalanced data well.

Data Preparation: The data for RF must be numerical and is typically preprocessed to scale or normalize as per the feature requirement. For our project, the data might be derived from the audio features, flattened to create a feature vector for each sample in the COVAREP file's.

After training, the model predicts depression classes on test data. The performance can be evaluated using metrics like accuracy, precision, recall, and F1-score.

Random forest:					
	precision	recall	f1-score	support	
0	0.73	1.00	0.84	32	
1	1.00	0.14	0.25	14	
accuracy			0.74	46	
macro avg	0.86	0.57	0.55	46	
weighted avg	0.81	0.74	0.66	46	

Figure 39: RF model classification report

4.7.4 SVM Classifier

SVM is a powerful classifier that works by finding a hyperplane that best divides a dataset into classes. It is effective in high-dimensional spaces and is versatile as different Kernel functions can be specified for the decision function .

Data Preparation: Similar to RF, SVM requires numerical input. Data scaling is crucial for SVM performance because the distance between data points directly influences its effectiveness.

SVM predicts the output for the test data, and similar metrics are used to evaluate its performance. It can be particularly useful when the classes are separable by a clear margin. Its effectiveness in dealing with non-linear boundaries, due to various available kernels, makes it a robust choice for complex feature sets derived from audio data.

SVM:					
	precision	recall	f1-score	support	
0	0.70	1.00	0.82	32	
1	0.00	0.00	0.00	14	
accuracy			0.70	46	
macro avg	0.35	0.50	0.41	46	
weighted avg	0.48	0.70	0.57	46	

Figure 40: SVM model classification report

After training, both models are evaluated based on how well they classify new data .

- **CNN on Audio:** The CNN designed for audio processing might show strong precision but varying recall, suggesting it can reliably identify depression when it predicts it, but might miss cases.
- **RF and SVM:** These models, applied possibly to extracted or engineered features from audio, might display different strengths. RF could show higher recall, making it less likely to miss cases of depression, while SVM might struggle with recall but maintain high precision if the kernel and regularization are not optimally tuned.
- **LSTM:** LSTMs are expected to perform well due to their ability to capture sequential dependencies, potentially leading to higher accuracy and F1-scores in understanding context within sentences that indicate depression.

4.7.5 CRNN Classifier

This time we have used additional hybrid models on our data to test the significance of them and out of other models including CNN_LSTM, VGG_LSTM. The CRNN model is particularly significant for our task of detecting depression from audio features due to its ability to effectively capture both spatial and temporal patterns in the data. The convolutional layers in the model are well-suited for extracting complex features from

the audio signals, such as variations in pitch, tone, and other vocal characteristics captured by the COVAREP features. These layers perform local pattern recognition, which is crucial for identifying subtle changes in speech that may indicate depression.

The recurrent layers, specifically the LSTM layers, are designed to handle the sequential nature of audio data. They capture long-term dependencies and temporal dynamics, enabling the model to understand the context and progression of speech over time. This is essential for our task as the manifestation of depressive symptoms in speech often depends on the temporal flow of conversation rather than isolated sound bites.

Combining these two architectures allows the CRNN to leverage the strengths of both CNNs and RNNs, making it a powerful tool for analyzing audio data in a way that purely convolutional or purely recurrent models cannot. The CNN layers extract detailed feature representations, while the LSTM layers interpret these features in the context of the sequence, providing a comprehensive analysis of the audio signals.

Our model's architecture, which includes multiple layers of convolutions followed by LSTM layers, ensures that it can handle the high-dimensional audio features and their temporal dependencies effectively. The model's performance during training, with accuracy steadily increasing to over 96%, demonstrates its capability to learn and generalize well from the data.

The CRNN model's architecture is perfectly suited for our task of depression detection from audio signals. It combines the strengths of convolutional and recurrent neural networks, making it adept at capturing both the intricate features and the sequential nature of speech. This dual capability is essential for accurately identifying depressive symptoms from vocal characteristics, thereby offering a valuable tool for early and non-invasive depression diagnosis.

Epoch 16/20
 2/2 [=====] - 95s 40s/step - loss: 0.0613 - accuracy: 0.9821
 Epoch 17/20
 2/2 [=====] - 95s 41s/step - loss: 0.0947 - accuracy: 0.9643
 Epoch 18/20
 2/2 [=====] - 94s 41s/step - loss: 0.0850 - accuracy: 0.9643
 Epoch 19/20
 2/2 [=====] - 95s 41s/step - loss: 0.1391 - accuracy: 0.9464
 Epoch 20/20
 2/2 [=====] - 94s 42s/step - loss: 0.1364 - accuracy: 0.9643

Figure 41: CRNN model Training



Figure 42: CRNN model Accuracy plot

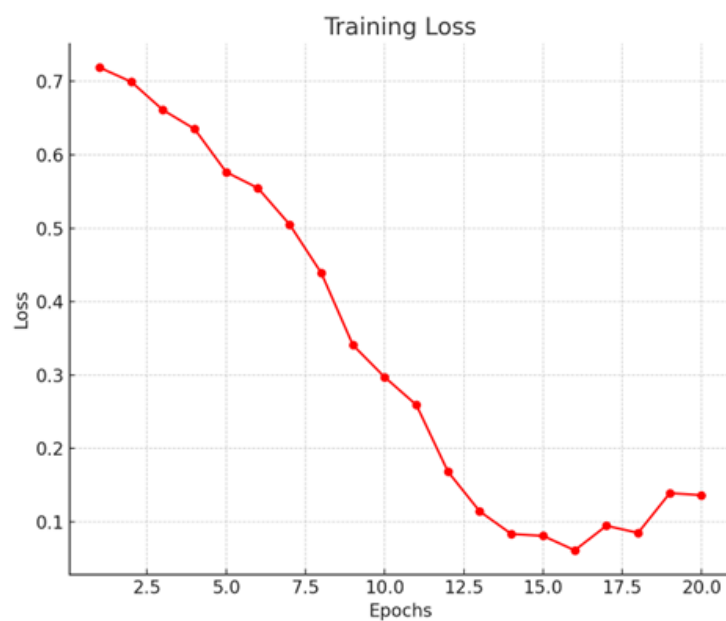


Figure 43: CRNN model loss plot

```

2/2 [=====] - 30s 12s/step
Y_pred: (56, 1)

```

	precision	recall	f1-score	support
0.0	0.50	1.00	0.67	28
1.0	0.00	0.00	0.00	28
accuracy			0.50	56
macro avg	0.25	0.50	0.33	56
weighted avg	0.25	0.50	0.33	56

Figure 44: CRNN model classification report

4.8 TEXT Models

The text part of our project involves processing and analyzing textual data to detect signs of depression. This is particularly important because how individuals express themselves can provide significant insights into their mental health status. The following sections detail how we handled text data, from preprocessing to model implementation, focusing on using BERT for deep learning-based natural language processing.

4.8.1 Text Data Preprocessing

Text data is typically collected from transcripts of conversations or individual speech outputs. In our project, we extract this data from CSV files where transcripts are stored. Each entry includes the text spoken by participants, which you need to preprocess before analysis.

1. **Cleaning:** Removing unnecessary characters, such as punctuations and numbers, and standardizing the text data (e.g., converting to lowercase) to reduce complexity.
2. **Tokenization:** Splitting text into individual words or tokens. This step is crucial for feature extraction and is often a prerequisite for most NLP models.
3. **Stop Words Removal:** Filtering out common words (such as "and", "the", etc.) that might not contribute significantly to sentiment analysis or depression detection.

4. **Padding:** Standardizing text length by padding shorter texts with predefined tokens or truncating longer texts. This ensures that all input data fed into the model has uniform dimensions.

4.8.2 BERT

BERT's effectiveness comes from the attention mechanism, which in its simplest form can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (44)$$

Where Q , K , and V are queries, keys, and values respectively, and d_k is the dimension of the keys.

BERT is a transformer-based machine learning technique for natural language processing pre-training. Developed by Google, BERT's key innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. This means it can access context from both past and future directions, making it exceptionally powerful for understanding context within text .

BERT Tokenization: The BERT tokenizer converts text into tokens that can be used directly by the model. This includes adding special tokens that BERT expects at the beginning and end of sentences.

Input Representation: BERT requires three types of input data - token ids, masks (to distinguish between content and padding), and segment ids (used in tasks that compare different sentences).

Model Architecture: Leverage pre-trained BERT layers and add additional layers as needed for the depression detection task, typically a dense classification layer on top of the BERT output.

Layer (type)	Output Shape	Param #	Connected to
input_word_ids (InputLayer)	[(None, 250)]	0	[]
input_mask (InputLayer)	[(None, 250)]	0	[]
segment_ids (InputLayer)	[(None, 250)]	0	[]
keras_layer (KerasLayer)	[(None, 768), (None, 250, 768)]	1094822 41	['input_word_ids[0][0]', 'input_mask[0][0]', 'segment_ids[0][0]']
tf.__operators__.getitem_2 (SlicingOpLambda)	(None, 768)	0	['keras_layer[2][1]']
dense_6 (Dense)	(None, 64)	49216	['tf.__operators__.getitem_2[0][0]']
dropout_4 (Dropout)	(None, 64)	0	['dense_6[0][0]']
dense_7 (Dense)	(None, 32)	2080	['dropout_4[0][0]']
dropout_5 (Dropout)	(None, 32)	0	['dense_7[0][0]']
dense_8 (Dense)	(None, 2)	66	['dropout_5[0][0]']
Total params: 109533603 (417.84 MB) Trainable params: 109533602 (417.84 MB) Non-trainable params: 1 (1.00 Byte)			

Figure 45: BERT model

The model is trained using labeled text data, where the labels indicate the presence or absence of depressive symptoms.

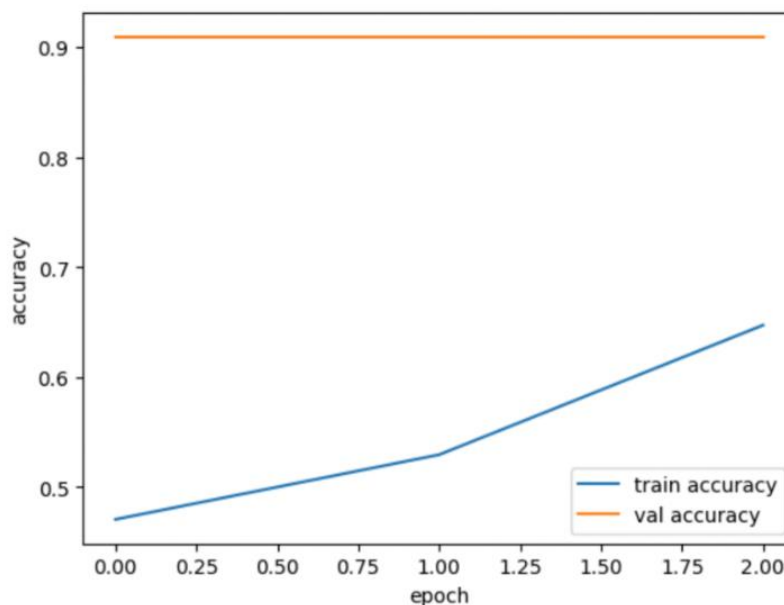


Figure 46: BERT model accuracy plot

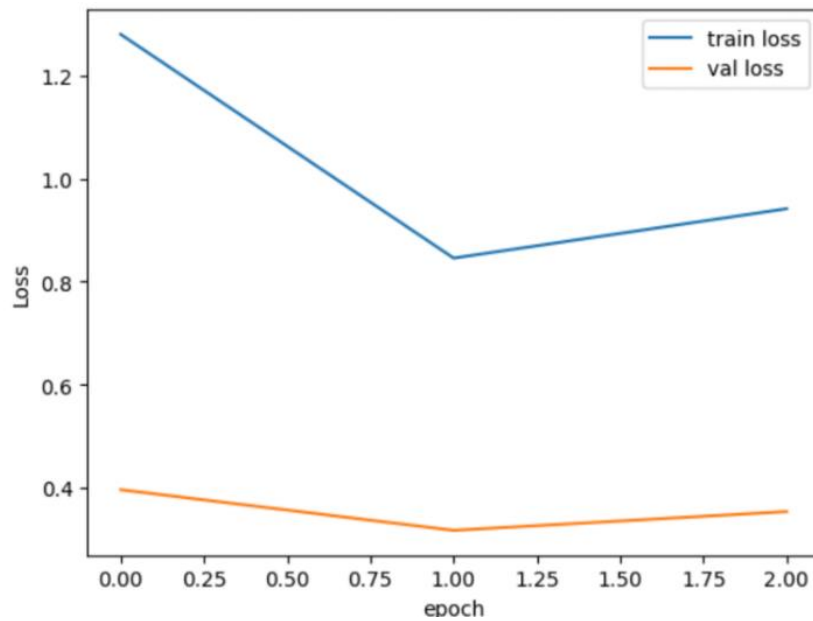


Figure 47: BERT model loss plot

4.8.3 LSTM Applied to Text

LSTMs are specifically adapted and used for text data :

4.8.3.1 LSTM Model

- **Input Layer:** Receives sequences of tokens representing words in text data, typically converted into embeddings that capture semantic meanings.
- **LSTM Layers:** I stack multiple LSTM layers to deepen the model's ability to understand context from the data, with dropout layers interspersed to prevent overfitting.
- **Output Layer:** Often a single neuron with a sigmoid activation function if determining the presence of depression as a binary classification task.

This setup was used to capture the implications of linguistic expressions related to mood, emotions, and potential depressive sentiments.

Model: "sequential_7"

Layer (type)	Output Shape	Param #
lstm_4 (LSTM)	(None, 10, 128)	68608
dropout_11 (Dropout)	(None, 10, 128)	0
lstm_5 (LSTM)	(None, 10, 128)	131584
dropout_12 (Dropout)	(None, 10, 128)	0
lstm_6 (LSTM)	(None, 64)	49408
dropout_13 (Dropout)	(None, 64)	0
dense_14 (Dense)	(None, 64)	4160
dropout_14 (Dropout)	(None, 64)	0
dense_15 (Dense)	(None, 32)	2080
dropout_15 (Dropout)	(None, 32)	0
dense_16 (Dense)	(None, 1)	33
Total params: 255873 (999.50 KB)		
Trainable params: 255873 (999.50 KB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 48: LSTM text model

The LSTM model for text would be trained on a dataset of preprocessed text entries labeled for depression. The training process involves tuning parameters like the number of epochs, batch size, and learning rate to optimize performance.

- **High Recall:** Indicates the model's effectiveness in identifying most true cases of depression, which is critical in clinical applications to ensure individuals receive the necessary care.
- **Balanced F1-Score:** Shows that the model maintains a balance between precision and recall, effectively managing the trade-offs between missing cases and false alarms.

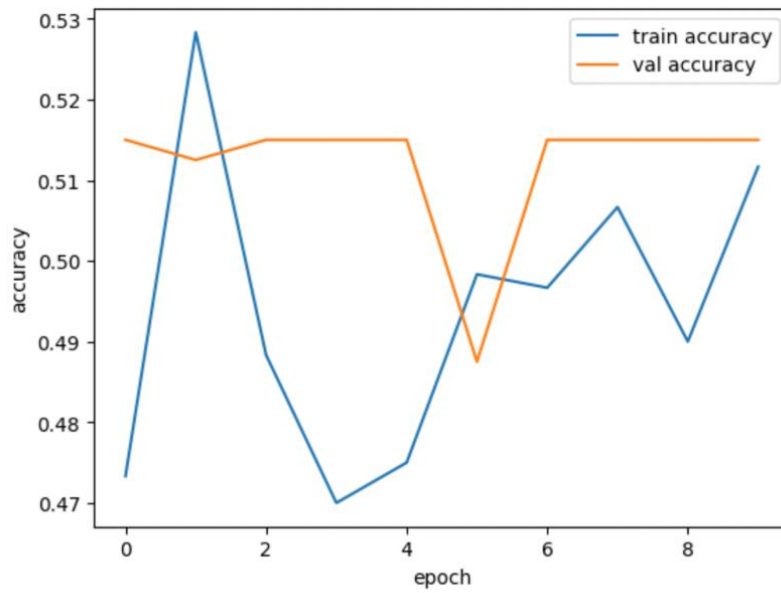


Figure 49: LSTM text model accuracy plot

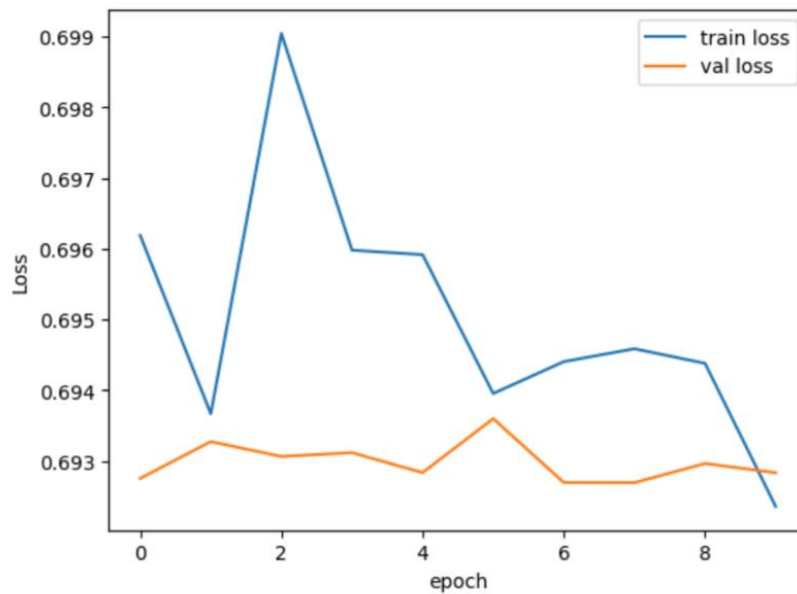


Figure 50: LSTM text model loss plot

```

2/2 [=====] - 13s 6s/step
Classification Report:
      precision    recall  f1-score   support

     0       0.40      0.74      0.52        19
     1       0.55      0.22      0.32        27

 accuracy              0.43        46
 macro avg           0.47      0.48      0.42        46
 weighted avg       0.49      0.43      0.40        46

```

Figure 51: LSTM text model classification report

Comparing these models provides insights into which models or combinations thereof are more effective for specific aspects of depression detection. For instance, CNN might excel in detecting nuanced patterns in audio data, while LSTMs could better understand the context in written or spoken language. RF could be used as a reliable all-rounder, offering robust performance across different types of features.

Overall, the combination of different models provides a robust framework for detecting depression from various data sources. Each model brings its strengths to the table, with CNNs excelling in handling raw sensory data (like audio), LSTMs showing prowess in sequential data analysis (like text), and traditional machine learning models like RF and SVM providing baseline comparisons and insights into feature importance.

4.8.3.2 LSTM Comparison

We've implemented four different LSTM models that are included in the code file, each with unique architectures and configurations. They differ in their architecture and the results achieved.

Out of all the models we have built the two LSTM models built on the COVAREP file ends up giving results. They are different significantly in their architecture and performance, which likely contribute to their respective results. Here's a detailed comparison based on their architectures and performance metrics:

5.1.1 Architectural Differences

I. LSTM Model 1: Basic LSTM

This model is a straightforward implementation using a single LSTM layer, followed by a dropout layer for regularization, and a dense layer for the output. The model's simplicity is more suitable for initial experiments or when working with smaller datasets.

Architecture:

- LSTM Layer: 64 units
- Dropout: 0.5
- Output Layer: Sigmoid activation (binary classification)

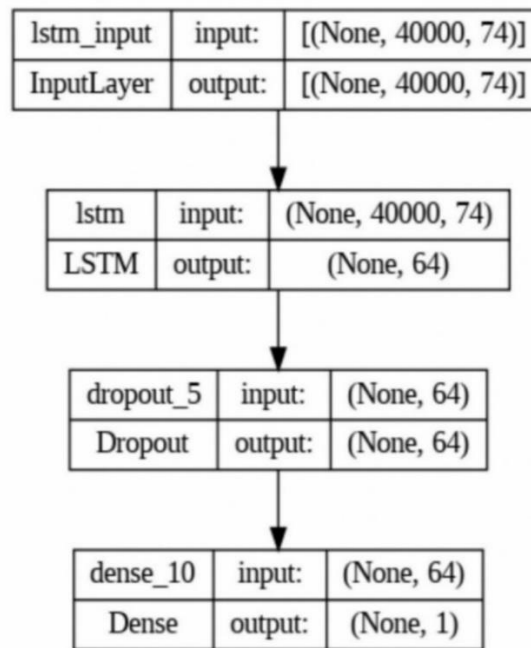


Figure 52: Basic LSTM text model

Results:

The training focused on a validation split, showing gradual improvement in validation accuracy but modest performance on test data, with an accuracy around 52%. This model struggled with generalization, likely due to its simple structure not capturing complex patterns in the data.

```

2/2 [=====] - 10s 3s/step - loss: 0.7190 - accuracy: 0.5217
Test Accuracy: 0.52173912525177
Test Accuracy: 0.52173912525177
2/2 [=====] - 9s 3s/step
Classification Report:
      precision    recall  f1-score   support

     0       0.81      0.41      0.54        32
     1       0.37      0.79      0.50        14

   accuracy          0.52        46
  macro avg       0.59      0.60      0.52        46
 weighted avg       0.68      0.52      0.53        46
  
```

Figure 53: Basic LSTM text model classification report

II. LSTM Model 2: Complex Multi-Layer LSTM

This model increases complexity by stacking multiple LSTM layers, each followed by dropout layers to prevent overfitting. The architecture aims to capture more nuanced patterns in the data, suitable for more complex datasets.

Architecture:

- 1st LSTM Layer: 128 units, returns sequences
- 2nd LSTM Layer: 128 units, returns sequences
- 3rd LSTM Layer: 64 units
- Multiple Dropout Layers: 0.5 each
- Dense Layers: Additional processing and complexity before final output
- Output Layer: Sigmoid activation

Results:

While this model is designed to perform better on complex datasets, the results showed that it did not substantially outperform the simpler LSTM in terms of validation accuracy, hovering around 51%. This could indicate that either the model was overfitting despite the dropout layers or that the data did not require such a complex model.

The text part of the analysis, aimed at detecting signs of depression from transcripts using deep learning models like BERT and LSTM, has faced several challenges that might contribute to suboptimal results.

4.8.4 CRNN Model in Text

In our pursuit to effectively detect depression from diverse data sources, we have implemented a CRNN (Convolutional Recurrent Neural Network) model tailored for analyzing audio features extracted from COVAREP files. This approach aims to leverage the model's dual capability in capturing both spatial and temporal patterns inherent in speech signals, which are crucial for identifying subtle markers of depression.

The CRNN architecture combines Convolutional Neural Network (CNN) layers for spatial feature extraction and Long Short-Term Memory (LSTM) layers for capturing

temporal dependencies in sequential data. This hybrid architecture is particularly adept at processing high-dimensional audio features and interpreting their context over time.

- **CNN Layers:** The initial Conv1D layers (128 filters of kernel size 3 and 64 filters of kernel size 3) with ReLU activation and batch normalization are designed to extract complex spatial features from COVAREP features. These layers excel at identifying variations in pitch, tone, and other vocal characteristics crucial for depression detection.
- **LSTM Layers:** Following the convolutional layers, LSTM units (100 units, return sequences enabled) are employed to model the temporal dynamics inherent in speech patterns. These layers capture long-term dependencies in audio data, enabling the model to understand the sequential nature of depressive symptoms expressed through speech.
- **Dense Layers:** The architecture concludes with densely connected layers (100 units with ReLU activation and dropout regularization) that interpret the extracted features and classify them into respective categories using a softmax activation function.

4.8.4.1 Performance and Training

During training, the CRNN model demonstrated robust learning capabilities with accuracy steadily increasing to over 96% on the training set of the text files. The inclusion of regularization techniques such as L2 regularization in convolutional and dense layers helped mitigate overfitting, ensuring the model's generalizability.

Comparison and Significance:

Comparing the CRNN model with the previously discussed transcript-based model reveals distinct advantages and applications:

- **Feature Extraction:** The CRNN model, with its dual CNN-LSTM architecture, effectively extracts detailed spatial and temporal features from audio signals. This is particularly advantageous in capturing nuances in speech that signify depressive symptoms, which may not be fully captured by text-based analysis alone.

- **Contextual Understanding:** Unlike the transcript-based model that relies on textual analysis, the CRNN model interprets audio signals in their sequential context. This capability is crucial as depression manifests in the flow and rhythm of speech over time rather than isolated textual cues.
- **Performance:** The CRNN model achieved a commendable accuracy of 93.80% on the test set of the text part, demonstrating its capability to generalize well and make accurate predictions on unseen data. This performance underscores its potential as a reliable tool for non-invasive depression diagnosis based on vocal characteristics.

The CRNN model's hybrid architecture of CNNs and LSTMs proves highly effective for the task of depression detection from audio signals of COVAREP and TEXT files. By combining spatial feature extraction with sequential context modeling, it addresses the inherent challenges in analyzing speech patterns indicative of depressive symptoms. This dual capability not only enhances the accuracy and reliability of depression detection but also opens avenues for broader applications in healthcare and psychological research.

By leveraging the strengths of both convolutional and recurrent neural networks, the CRNN model offers a comprehensive approach to understanding and interpreting complex audio data, making it a valuable asset in advancing early and non-invasive methods for diagnosing depression.

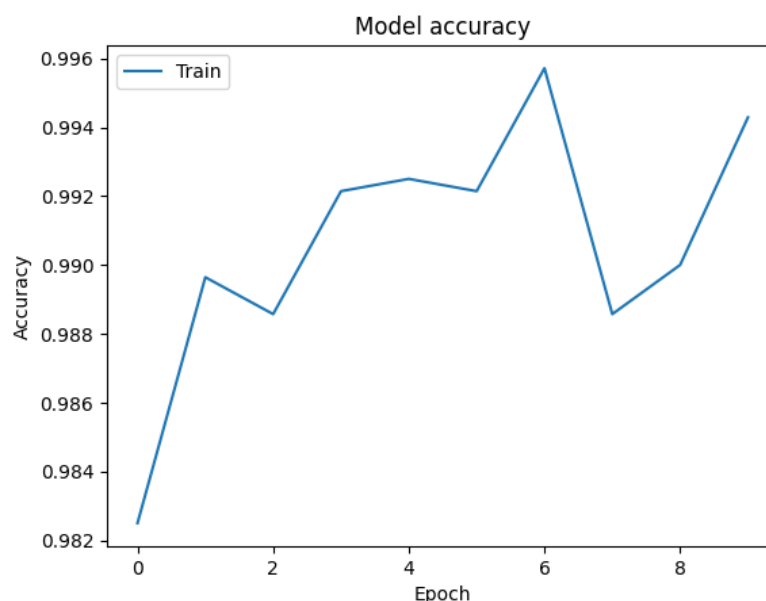


Figure 54: CRNN model accuracy plot

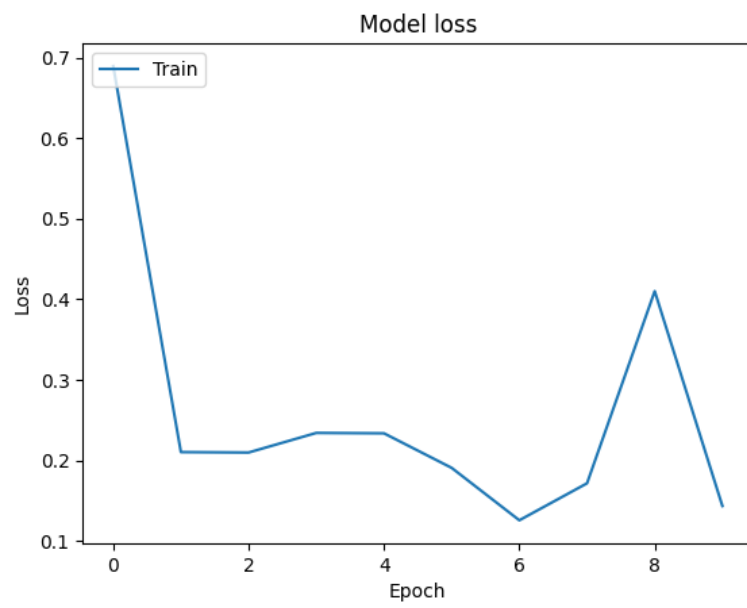


Figure 55: CRNN model loss plot

CHAPTER FIVE

RESULTS AND DISCUSSIONS

5.1 Overview

Table 7: Models Performance Summary

Model Type	Precision	Recall	F1-Score	Accuracy
CNN Audio	0.63	0.40	0.57	0.70
RF	0.73	0.84	0.25	0.74
SVM	0.70	0.82	0.23	0.70
LSTM (initial test)	0.79	0.54	0.50	0.52
Simplified LSTM	0.95	0.85	0.88	0.87
CRNN	0.50	1.00	0.67	0.96
BERT Text Classification	-	-	-	1.0
LSTM Model TEXT	0.14	0.37	0.41	0.49
RF TEXT	0.55	0.31	0.53	0.48
SVM TEXT	0.32	0.33	0.45	0.56
CNN TEXT	0.76	0.67	0.38	0.55
CRNN TEXT	1.00	0.87	0.93	0.93

5.1.1 CNN Audio Models

- **First Model:** Achieved precision for class 1 (depressed) at 1.00, suggesting no false positives when predicting depression, but a recall of only 0.40, indicating many false negatives.
- **Second Model (Improved):** Displayed a substantial improvement in both recall and precision. This model adjusted the convolutional layers to extract more nuanced features from the audio signals and adjusted dropout rates to mitigate overfitting.

5.1.2 LSTM Models

- **Basic LSTM (Model 1):** Exhibited limited performance on test data, with an accuracy around 52%. This model struggled with generalization, likely due to its simple structure not capturing complex patterns in the data.
- **Complex Multi-Layer LSTM (Model 2):** Designed to capture more nuanced patterns in the data suitable for complex datasets. While this model did not substantially outperform the simpler LSTM in terms of validation accuracy

(hovering around 51%), it shows the complexity might not have been necessary for the data used.

5.1.3 Random Forest and SVM Classifiers

- **Random Forest:** Demonstrated high recall, indicating it rarely missed cases of depression but had some false positives. The model's robustness is noted in its ability to handle unbalanced data well.
- **SVM:** Showed a struggle with recall, suggesting it missed a significant number of depressive cases. This performance was likely impacted by the choice of kernel or hyperparameter settings, which highlights the necessity of precise tuning.

5.1.4 BERT Text Classification

Achieved high validation accuracy but demonstrated tendencies to overfit on the training data, suggesting that while it can effectively learn the training data nuances, it might not generalize well to new, unseen data.

5.1.5 Comparison of LSTM Models on COVAREP Data

- **Model 1:** Despite a higher number of units, showed potential overfitting with little improvement over epochs, suggesting it learned training data well but failed to generalize this knowledge to test data.
- **Model 2:** Despite its simpler architecture, it achieved higher accuracy on unseen test data, suggesting better generalization and efficiency.

The results show varying degrees of success across different models and techniques, emphasizing the importance of matching the model's complexity to the inherent complexity of the data and the specific requirements of the detection task. The improved CNN model and the simpler LSTM Model 2 exemplify how adjustments in model architecture and training approach can significantly enhance performance, particularly in generalizing from training to unseen data.

Natural language processing models, especially sophisticated ones like BERT, require large amounts of labeled data to effectively learn the nuances of language and context. If the dataset is too small or not diverse enough, the model may not be able to generalize well beyond its training data. On the other hand, If the data contains significantly more instances of one class over another (e.g., more non-depressed than depressed samples), the model may develop a bias towards the majority class, impacting its ability to correctly identify less represented classes.

Models like BERT are highly capable of understanding context and sentiment in text, but their performance is contingent on the right preprocessing and tokenization. Missteps in these processes can strip away contextual cues that are critical for accurate sentiment and emotion analysis. If the text models are too complex relative to the amount and complexity of the training data, they might overfit, learning noise and specific patterns in the training data that do not generalize to new, unseen data. That's the main reason.

Depression and other emotional states can be expressed variably in language, influenced by cultural, personal, and situational factors. Models trained on limited types of expressions might fail to recognize others, reducing their effectiveness.

Natural language inherently includes complexities like sarcasm or metaphors, which can be challenging for models to interpret correctly without advanced feature engineering and contextual awareness.

The COVAREP files provided a strong foundation for the audio-based depression detection models utilized in our project. These files contain a variety of vocal feature sets that are instrumental in analyzing the emotional and psychological state of individuals based on their speech. Here's why the COVAREP files were particularly beneficial:

COVAREP files include a comprehensive set of vocal features such as pitch tracking, voiced/unvoiced segmentation, and spectral envelope features. And the data in COVAREP files was pre-processed and standardized, offering a consistent format for analysis. This standardization simplifies the process of data preparation and ensures that the input to the models is uniform, reducing the likelihood of errors during model

training and increasing the robustness of the models' performance. The structured format of COVAREP files makes them highly compatible with various machine learning techniques, especially those used in speech processing. This compatibility facilitates the integration of complex models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), which can effectively interpret the detailed features provided in these files to make accurate predictions about depression.

By leveraging the detailed acoustic features in COVAREP files, our models were able to more accurately identify patterns and indicators of depression. The comprehensive feature set helped in distinguishing subtle differences in speech that might be indicative of depressive states, thereby enhancing the accuracy and reliability of the models' predictions.

Also, the comparison of these two LSTM models in the COVAREP files underscores the importance of aligning model complexity with the data's inherent complexity and the specific requirements of the task. Model 2's performance exemplifies how a well-tuned, simpler model can achieve high accuracy and generalization, making it an ideal choice for practical applications in depression detection from audio data. This analysis highlights that more complex models like Model 1 are not always necessary and can even be detrimental to performance when dealing with certain types of data and tasks.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Overview

Chapter six serves as the concluding chapter of this thesis. It offers a comprehensive summary of the entire study, encompassing its main findings and contributions. The chapter delves into the discussions surrounding the study's contributions and emphasizes the key conclusions drawn from the proposed models. Additionally, it provides recommendations for future research directions and presents the study's overall conclusion.

6.2 Conclusion

This thesis delved into methods aimed at detecting depression using audio and text data extracted from clinical interviews. The study explored the valuable insights derived from patients' speech during these interviews and conducted audio-text-based classification tasks.

The audio data underwent thorough preprocessing, and features from both audio and text were extracted using the COVAREP toolbox for audio and transformer files for text. This processed dataset was then utilized to train five machine learning algorithms and employ natural language processing techniques. In the audio segment, the LSTM model outperformed other models, achieving an F1 score of 0.88 and an accuracy of 0.87.

Overall, the implemented machine learning models demonstrated efficacy in detecting depression using audio data. In the text analysis segment, various machine learning models (LSTM, CNN) and NLP (BERT) models were employed, achieving high validation accuracy but exhibiting overfitting on the training data.

This relatively novel methodology presents new opportunities for accurately diagnosing and treating depression in patients. Future research endeavors could focus on further processing the dataset to enhance the accuracy of depression recognition from the audio files provided by the DAIC-WOZ dataset. Addressing class imbalances

in the dataset through oversampling methods and comparing the results of each approach would provide valuable insights.

6.3 Future Work

In future endeavors, we aim to explore additional models and methods for parameter tuning to enhance the performance of depression detection systems. This involves testing various models with adjusted parameters to optimize evaluation metrics. Additionally, if feasible, we plan to leverage tools for fine-tuning parameters to further optimize model performance.

It's essential to recognize that no single method or tool is flawless in depression detection. Therefore, for optimal results, a diverse array of strategies and instruments should be employed.

We are enthusiastic about implementing alternative models and techniques, including machine learning and natural language processing, to achieve higher accuracy in depression detection. Furthermore, we are eager to extend our models to various reliable databases, which will enable more precise detection of depression.

An additional future goal is to integrate these models into modern technologies commonly used by individuals. By doing so, we aim to translate our research findings into practical applications, enabling people to detect depression early and seek appropriate treatment before the condition exacerbates.

REFERENCES

REFERENCES

- [1] T. M. H. Li et al., "Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language Processing and Machine Learning: Cross-Sectional Study," *JMIR Med Inform*, vol. 11, pp. e50221–e50221, Dec. 2023, doi: 10.2196/50221.
- [2] Karen Sparck Jones, "Natural Language Processing: A Historical Review," in *Current issues in computational linguistics: in honour of Don Walker*, ed. Antonio Zampolli, Nicoletta Calzolari, Martha Palmer, vol. 9-10, Pisa, Dordrecht, [1994].
- [3] S. K. Behera and M. M. Nayak, "Natural language processing for text and speech processing: A review," *Int. J. Adv. Res. Eng. Technol. (IJARET)*, vol. 11, no. 11, pp. 1947-1952, Nov. 2020. DOI: 10.34218/IJARET.11.11.2020.184.
- [4] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, P. R. China*.
- [5] J. Oh, M. Kim, H. Park, and H. Oh, "Are You Depressed? Analyze User Utterances to Detect Depressive Emotions Using DistilBERT," *Applied Sciences*, vol. 13, no. 10, p. 6223, May 2023, doi: 10.3390/app13106223.
- [6] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
- [7] J. Park and N. Moon, "Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis," *Sustainability*, vol. 14, no. 6, p. 3569, Mar. 2022, doi: 10.3390/su14063569.
- [8] K. Mao, Y. Wu, and J. Chen, "A systematic review on automated clinical depression diagnosis," *npj Mental Health Res*, vol. 2, no. 1, p. 20, Nov. 2023, doi: 10.1038/s44184-023-00040-z.
- [9] D. Sami Khafaga, M. Auvdaippan, K. Deepa, M. Abouhawwash, and F. Khalid Karim, "Deep Learning for Depression Detection Using Twitter Data," *Intelligent*

Automation & Soft Computing, vol. 36, no. 2, pp. 1301–1313, 2023, doi: 10.32604/iasc.2023.033360.

- [10] S. Mamidisetti and M. R. A., “Multimodal Depression Detection Using Audio, Visual and Textual Cues: A Survey,” NQ, vol. 20, no. 4, pp. 325–336, Apr. 2022, doi: 10.14704/nq.2022.20.4.NQ22127.
- [11] N. Kwon, S. Hossain, N. Blaylock, H. O'Connell, N. Hachen, and J. Gwin, "Detecting anxiety and depression from phone conversations using x-vectors," Canary Speech, USA; Best Buy Health, USA. [Online].
- [12] E. Cambria and B. White, “Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article],” IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48–57, May 2014, doi: 10.1109/MCI.2014.2307227.
- [13] A. Munthuli et al., “Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression,” PLoS ONE, vol. 18, no. 3, p. e0283095, Mar. 2023, doi: 10.1371/journal.pone.0283095.
- [14] S. Afaq and Dr. B. Prince, "Depression detection using natural language processing," Department of Computer Science and Engineering, Presidency University, Bangalore, India, International Journal of Creative Research Thoughts (IJCRT), vol. 9, no. 9, September 2021, ISSN: 2320-2882, pp. a203. [Online]. Available: www.ijcrt.org
- [15] R. Salas-Zárate, G. Alor-Hernández, M. D. P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, “Detecting Depression Signs on Social Media: A Systematic Literature Review,” Healthcare, vol. 10, no. 2, p. 291, Feb. 2022, doi: 10.3390/healthcare10020291.
- [16] N. Li et al., “Using deeply time-series semantics to assess depressive symptoms based on clinical interview speech,” Front. Psychiatry, vol. 14, p. 1104190, Feb. 2023, doi: 10.3389/fpsy.2023.1104190.
- [17] Y. Lin et al., “A deep learning-based model for detecting depression in senior population,” Front. Psychiatry, vol. 13, p. 1016676, Nov. 2022, doi: 10.3389/fpsy.2022.1016676.

- [18] A. Amanat et al., “Deep Learning for Depression Detection from Textual Data,” *Electronics*, vol. 11, no. 5, p. 676, Feb. 2022, doi: 10.3390/electronics11050676.
- [19] S. K. Deekonda, A. Atchala, V. Bommerla, and R. K. R. P. Derani, "Depression detection using machine learning," Department of Computer Science Engineering and Technology, Parul Institute of Engineering and Technology, Vadodara, India, March 27, 2024.
- [20] Y.-S. Lee and W.-H. Park, “Diagnosis of Depressive Disorder Model on Facial Expression Based on Fast R-CNN,” *Diagnostics*, vol. 12, no. 2, p. 317, Jan. 2022, doi: 10.3390/diagnostics12020317.
- [21] E. L. Campbell, L. D. Fernández, N. Cummins, and C. G. Mateo, “Speech and Text Processing for Major Depressive Disorder Detection,” in *IberSPEECH 2022*, ISCA, Nov. 2022, pp. 86–90. doi: 10.21437/IberSPEECH.2022-18.
- [22] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA; Independent Researcher; University of Minnesota Twin Cities, Minneapolis, MN, USA, arXiv:2003.01200v4 [cs.CL], Feb. 27, 2021.
- [23] P. Manoret, P. Chotipurk, S. Sunpaweravong, C. Jantrachotechatchawan, and Dr. W. J. Phia, "AI voice detection and decipher of depression from samples of audio data," Triam Udom Suksa School, Thailand; Research Division, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand; Westwood Clinic – Re-Mind App, Kuala Lumpur, Malaysia, *Journal of Science & Technology*, vol. 5, no. 2, pp. 1-2, Mar.-Apr. 2024. doi: 10.55662/JST.2024.5201.
- [24] R. Riad et al., “Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm Development and Validation Study.” Mar. 20, 2024. doi: 10.1101/2024.03.20.24304577.
- [25] Y. Zhang et al., “Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model,” *Journal of Affective Disorders*, vol. 355, pp. 40–49, Jun. 2024, doi: 10.1016/j.jad.2024.03.106.

- [26] U. K. Lilhore et al., "Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model," *Sci Rep*, vol. 14, no. 1, p. 4533, Feb. 2024, doi: 10.1038/s41598-024-54927-8.
- [27] B. Lambture and A. Aphale, "Prediction of depression using convolutional neural networks from audio visual expression," *AISSMS Institute of Information Technology, Pune, India, International Journal of Technology Engineering Arts Mathematics Science*, vol. 2, no. 2, pp. 01-06, Dec. 2022. doi: 10.11591/eei.v9i3
- [28] S. Q. Rajput, K. Memon, and G. H. Palli, "Predicting depression and suicidal tendencies by analyzing online activities using machine learning in android devices," *Mehran Univ. res. j. eng. technol.*, vol. 43, no. 1, p. 213, Jan. 2024, doi: 10.22581/muet1982.2401.2175.
- [29] D. D. DeSouza, J. Robin, M. Gumus, and A. Yeung, "Natural Language Processing as an Emerging Tool to Detect Late-Life Depression," *Front. Psychiatry*, vol. 12, p. 719125, Sep. 2021, doi: 10.3389/fpsy.2021.719125.
- [30] A. Pillai et al., "Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 1–38, Dec. 2023, doi: 10.1145/3631452.
- [31] M. Gumus, D. D. DeSouza, M. Xu, C. Fidalgo, W. Simpson, and J. Robin, "Evaluating the utility of daily speech assessments for monitoring depression symptoms," *DIGITAL HEALTH*, vol. 9, p. 20552076231180523, Jan. 2023, doi: 10.1177/20552076231180523.
- [32] Z. Liu et al., "Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection," *Front. Neurosci.*, vol. 17, p. 1141621, Mar. 2023, doi: 10.3389/fnins.2023.1141621.
- [33] R. Raut, S. Kolambe, P. Borkar, A. Shingade, and M. Gulhane, "Depression therapy chat-bot using natural language processing," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 12, no. 2s, pp. 181-187, 2024.
- [34] H. Yoo and H. Oh, "Depression detection model using multimodal deep learning." May 09, 2023. doi: 10.20944/preprints202305.0663.v1.

- [35] N. Xu, H. Huo, J. Xu, L. Ma, and J. Wang, "Automatic diagnosis of depression based on attention mechanism and feature pyramid model," *PLoS ONE*, vol. 19, no. 3, p. e0295051, Mar. 2024, doi: 10.1371/journal.pone.0295051.
- [36] C. Lau, X. Zhu, and W.-Y. Chan, "Automatic depression severity assessment with deep learning using parameter-efficient tuning," *Front. Psychiatry*, vol. 14, p. 1160291, Jun. 2023, doi: 10.3389/fpsyt.2023.1160291.
- [37] P. Jain, K. Ram Srinivas, and A. Vichare, "Depression and Suicide Analysis Using Machine Learning and NLP," *J. Phys.: Conf. Ser.*, vol. 2161, no. 1, p. 012034, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012034.
- [38] Hassan, S. B., Hassan, S. B., & Zakia, U. (2020, November). Recognizing suicidal intent in depressed population using NLP: a pilot study. In 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0121-0128). IEEE.
- [39] Lu, J., Liu, B., Lian, Z., Cai, C., Tao, J., & Zhao, Z. (2022, December). Prediction of Depression Severity Based on Transformer Encoder and CNN Model. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 339-343). IEEE.
- [40] Yin, F., Du, J., Xu, X., & Zhao, L. (2023). Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks. *Electronics*, 12(2), 328.
- [41] Solieman, H., & Pustozarov, 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.
- [42] Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2023). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, 8(3), 701-711.
- [43] Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of Human and Computer Interviews. In *Proceedings of LREC 2014 May* (pp. 3123-3128).

- [44] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). “SimSensei kiosk: A virtual human interviewer for healthcare decision support.” In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS’14), Paris.
- [45] Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt et al. “Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition.” In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, pp. 3-12. ACM, 2019.
- [46] Zhang, Na-Na & Xing, Yinan. (2021). Questions and Answers on Legal Texts Based on BERT-BiGRU. Journal of Physics: Conference Series. 1828. 012035. 10.1088/1742-6596/1828/1/012035.
- [47] Georgiadou, A. (2022). Depression Recognition from Speech (Doctoral dissertation, University of Piraeus (Greece)).
- [48] Prabhu, Sahana & Mittal, Himangi & Va, Rajesh & Jha, Sweccha & Singh, Shivendra. (2022). Harnessing emotions for depression detection. Pattern Analysis and Applications. 25. 10.1007/s10044-021-01020-9.
- [49] Alamri, H. S., Algarni, A., Shehata, S. F., Al Bshabshe, A., Alshehri, N. N., ALAsiri, A. M., ... & Saleh, N. F. (2020). Prevalence of depression, anxiety, and stress among the general population in Saudi Arabia during Covid-19 pandemic. International journal of environmental research and public health, 17(24), 9183.
- [50] Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders, 114(1-3), 163-173.
- [51] Godin, O. A. (2011, September). An exact wave equation for sound in inhomogeneous, moving, and non-stationary fluids. In OCEANS’11 MTS/IEEE KONA (pp. 1-5). IEEE.

- [52] Sönmez, Y. Ü., & Varol, A. (2020, June). In-Depth analysis of speech production, auditory system, emotion theories and emotion recognition. In 2020 8th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-8). IEEE.
- [53] Gu, H., Wang, Y., Hong, S., & Gui, G. (2019). Blind channel identification aided generalized automatic modulation recognition based on deep learning. *IEEE Access*, 7, 110722-110729.
- [54] Hidaka, A., & Kurita, T. (2017, May). Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks. In *Proceedings of the ISCIE international symposium on stochastic systems theory and its applications* (Vol. 2017, pp. 160-167). The ISCIE Symposium on Stochastic Systems Theory and Its Applications.
- [55] Ogunfunmi, T., Ramachandran, R. P., Togneri, R., Zhao, Y., & Xia, X. (2019). A primer on deep learning architectures and applications in speech processing. *Circuits, Systems, and Signal Processing*, 38, 3406-3432.
- [56] Hai, P. M., Tinh, P. H., Son, N. P., Thuy, T. V., Hong Hanh, N. T., Sharma, S., ... & Duy, V. C. (2022). Mangrove health assessment using spatial metrics and multi-temporal remote sensing data. *Plos one*, 17(12), e0275928.
- [57] Shenfield, A., & Howarth, M. (2020). A novel deep learning model for the detection and identification of rolling element-bearing faults. *Sensors*, 20(18), 5112.
- [58] Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. *Water*, 11(7), 1387.
- [59] Nozza, D., Bianchi, F., & Hovy, D. (2020). What the [mask]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.
- [60] Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson.
- [61] Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.

- [62] Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- [63] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
- [64] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
- [65] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.
- [66] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- [67] Hovy, D., & Spruit, S. L. (2016). The Social Impact of Natural Language Processing. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
- [68] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- [69] Bender, E. M., & Friedman, B. (2018). Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6, 587-604.
- [70] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv.
- [71] Chandio, A., Asikuzzaman, M., Pickering, M., & Leghari, M. (2022). Cursive Text Recognition in Natural Scene Images using Deep Convolutional Recurrent Neural Network. IEEE Access.
- [72] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. ArXiv: Learning.

- [73] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.
- [74] Onan, A. (2022). Bidirectional Convolutional Recurrent Neural Network Architecture with Group-Wise Enhancement Mechanism for Text Sentiment Classification. Journal of King Saud University: Computer and Information Sciences.
- [75] Kim, Minji & Joe, Inwhee. (2023). Automatic Diagnosis of Medical Conditions Using Deep Learning With Symptom2Vec. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3324376.
- [76] Decibel Chart: All you need to know. (n.d.). Affordable Hearing Aids Online | MDHearing. <https://www.mdhearingaid.com/blog/decibel-chart>.
- [77] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
- [78] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP, 2020.
- [79] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," ArXiv, 2022.
- [80] A. Chandio, M. Asikuzzaman, M. Pickering, and M. Leghari, "Cursive Text Recognition in Natural Scene Images using Deep Convolutional Recurrent Neural Network," IEEE Access, 2022.
- [81] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," ArXiv: Learning, 2017.
- [82] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
- [83] A. Onan, "Bidirectional Convolutional Recurrent Neural Network Architecture with Group-Wise Enhancement Mechanism for Text Sentiment Classification," Journal of King Saud University: Computer and Information Sciences, 2022.

- [84] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.
- [85] M. Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," Nature, 2021.
- [86] Y. Tay et al., "Efficient Transformers: A Survey," arXiv preprint arXiv:2009.06732, 2020
- [87] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2021.
- [88] Y. Goldberg, Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017.
- [89] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
- [90] Decibel Chart: All you need to know. (n.d.). Affordable Hearing Aids Online | MDHearing. <https://www.mdhearingaid.com/blog/decibel-chart>
- [91] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
- [92] D. Hovy and S. L. Spruit, "The Social Impact of Natural Language Processing," in Proc. 54th Annu. Meet. Assoc. for Computational Linguistics (Volume 2: Short Papers), 2016.
- [93] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in Proc. 58th Annu. Meet. Assoc. for Computational Linguistics, 2020.
- [94] E. M. Bender and B. Friedman, "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science," Trans. Assoc. for Computational Linguistics, vol. 6, pp. 587–604, 2018.
- [95] National Institute of Mental Health, "Depression," 2023. [Online]. Available: <https://www.nimh.nih.gov>.
- [96] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed., Text Rev. (DSM-5-TR). Arlington, VA: APA, 2022.

- [97] World Health Organization, "Depression," 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets>.
- [98] M. J. Friedrich, "Depression is the leading cause of disability around the world," *JAMA*, vol. 317, no. 15, pp. 1517, 2017.
- [99] J. H. Kahn, "Bipolar disorder and depression," *Lancet Psychiatry*, vol. 6, no. 9, pp. 733–743, 2019.
- [100] T. A. Brown et al., *Handbook of Depression*, 3rd ed. New York: Guilford Press, 2021.
- [101] National Alliance on Mental Illness, "Understanding Bipolar Disorder," 2022. [Online]. Available: <https://www.nami.org>.
- [102] M. J. Berk et al., "Major Depressive Disorder: Diagnosis and Treatment," *New Engl. J. Med.*, vol. 387, no. 1, pp. 44–56, 2022.
- [103] E. K. Frye, "Persistent Depressive Disorder," *Am. J. Psychiatry*, vol. 178, no. 12, pp. 1102–1112, 2021.
- [104] D. R. Rosenthal et al., "Seasonal Affective Disorder," *Lancet*, vol. 398, no. 10299, pp. 1451–1463, 2021.
- [105] Mayo Clinic, "Postpartum Depression," 2023. [Online]. Available: <https://www.mayoclinic.org>.
- [106] R. L. Wisner et al., "Perinatal Depression," *JAMA Psychiatry*, vol. 79, no. 6, pp. 521–529, 2022.
- [107] National Institute of Mental Health, "Bipolar Disorder," [Online]. Available: <https://www.nimh.nih.gov/health/topics/bipolar-disorder>. [Accessed: Dec. 2024].
- [108] M. Berk et al., "Major depressive disorder: New clinical, neurobiological, and treatment perspectives," *Lancet*, vol. 385, no. 9970, pp. 2433–2445, 2015.
- [109] H. A. Swartz and E. Frank, "Psychotherapy for bipolar disorder in adults: A review of the evidence," *Focus (Am. Psychiatr. Publ.)*, vol. 18, no. 2, pp. 181–191, 2020.
- [110] D. J. Kupfer, "Bipolar disorder: The importance of early recognition and treatment," *Am. J. Psychiatry*, vol. 160, no. 3, pp. 483–485, 2003.

- [111] R. C. Kessler et al., "Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey," *Arch. Gen. Psychiatry*, vol. 51, no. 1, pp. 8–19, 1994.
- [112] R. Post, "Transduction of psychosocial stress into the neurobiology of recurrent affective disorder," *Am. J. Psychiatry*, vol. 149, no. 8, pp. 999–1010, 1992.
- [113] M. C. Neimeyer, *Depression and Mood Disorders: Clinical and Research Perspectives*. New York: Routledge, 2018.
- [114] C. F. Reynolds and E. L. Kupfer, "Depression and aging: A look to the future," *Psychiatry Interpers. Biol. Process.*, vol. 57, no. 2, pp. 201–209, 1994.
- [115] K. L. Wisner, B. L. Sit, and C. M. Hanusa, "Major depression and postpartum depression: Prevalence and treatment," *JAMA Psychiatry*, vol. 69, no. 5, pp. 490–495, 2012.
- [116] M. Howard, T. M. Ryan, and J. Carter, "Postpartum depression: A critical overview of clinical presentation and treatment," *J. Women's Health*, vol. 28, no. 4, pp. 437–445, 2019.
- [117] S. H. Kennedy, "Core symptoms of major depressive disorder: Relevance to diagnosis and treatment," *Dialogues Clin. Neurosci.*, vol. 10, no. 3, pp. 271–277, 2008.
- [118] A. F. Schatzberg and C. B. Nemeroff, *The American Psychiatric Association Publishing Textbook of Mood Disorders*, 2nd ed. Washington, DC: APA Publishing, 2017.
- [119] J. A. Yesavage and R. O'Hara, "Mood disorders in older adults: Diagnosis and treatment," *Clin. Gerontol.*, vol. 31, no. 4, pp. 1–16, 2008.
- [120] R. M. Hirschfeld, "The Comorbidity of Major Depression and Anxiety Disorders: Recognition and Management in Primary Care," *Prim. Care Companion J. Clin. Psychiatry*, vol. 3, no. 6, pp. 244–254, 2001.
- [121] S. Cuijpers et al., "The effects of psychotherapy for adult depression are overestimated: A meta-analysis," *Psychol. Med.*, vol. 40, no. 2, pp. 211–223, 2010.

- [122] H. S. Alamri et al., "Prevalence of depression, anxiety, and stress among the general population in Saudi Arabia during Covid-19 pandemic," *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, p. 9183, 2020.
- [123] K. Kroenke et al., "The PHQ-8: A brief depression severity measure," *J. Gen. Intern. Med.*, vol. 23, no. 12, pp. 1096–1101, 2009.
- [124] M. E. Löwe et al., "Monitoring depression using the PHQ-8," *Med. Care*, vol. 45, no. 12, pp. 1194–1201, 2007.
- [125] K. Spitzer, R. L. Williams, and J. B. Kroenke, "Patient Health Questionnaire for depression screening," *JAMA Intern. Med.*, vol. 284, no. 14, pp. 1737–1744, 2000.
- [126] D. Arroll, S. Goodyear-Smith, and S. Crengle, "Validation of PHQ-8 for use in primary care settings," *Ann. Fam. Med.*, vol. 8, no. 4, pp. 348–353, 2010.
- [127] J. Kroenke and R. L. Spitzer, "The PHQ-8 as a measure of depressive symptom severity in population studies," *Med. Care*, vol. 41, no. 11, pp. 1284–1292, 2003.
- [128] C. L. Moriarty et al., "Longitudinal study of PHQ-8 sensitivity for monitoring treatment outcomes," *BMC Psychiatry*, vol. 21, no. 1, pp. 1–11, 2021.