DEPRESSION DETECTION USING DEEP LEARNING MODELS BASED ON MULTIMODAL SOCIAL MEDIA CONTENT

A Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

in

Computer Science & Engineering

by

Pavi Saraswat
(2K21/PHDCO/05)

Under the Supervision of

Dr. Rohit Beniwal
(Supervisor)
Department of Computer
Science & Engineering
Delhi Technological University

Department of Computer Science & Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

November, 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I certify that the dissertation titled "Depression Detection Using Deep Learning Models Based on Multimodal Social Media Content" submitted for the Doctor of Philosophy degree is my work and has not been submitted for the award of any degree or diploma to any other University or Institute. The work done in the thesis is original and has been done by me under the supervision of my supervisors.

I also mention that the research work is original and has not been submitted by me, in part or completely, to any other University or Institution for the award of any degree or diploma.

Pavi Saraswat

(Ph.D. Research Scholar)

Department of Computer Science and Engineering,

Delhi Technological University, Delhi

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE

This is to certify that the work contained in the thesis entitled "Depression Detection Using Deep Learning Models Based on Multimodal Social Media Content" submitted by Pavi Saraswat (2K21/PHDCO/05) for the award of the degree of Doctor of Philosophy to Delhi Technological University, India contains original research work carried out by her under my supervision.

She has fulfilled all the requirements as per the required standard for the submission of the thesis. I hereby confirm the originality of the work and certify that the thesis has not been submitted for the award of any degree or diploma at this or any other institution.

Dr. Rohit Beniwal

(Supervisor)

Assistant Professor

Department of Computer Science and Engineering

Delhi Technological University

ACKNOWLEDGMENT

I express my profound gratitude to Almighty God for providing me the strength, resilience, and guidance to pursue and complete this research journey. I am deeply indebted to my supervisor, **Dr. Rohit Beniwal**, for his invaluable mentorship, constant encouragement, and insightful suggestions throughout this journey. Dr. Beniwal's technical expertise and thoughtful guidance have been instrumental in overcoming the challenges faced during my research and he has been a constant source of motivation and support. His leadership and vision inspired me to strive for excellence. Also, my sincere thanks to **Prof. Manoj Kumar**, Professor and Head of the Department, Department of Computer Science and Engineering, for insightful comments and valuable suggestions. My sincere thanks also go to **Prof. Rahul Katarya**, DRC Chairperson, Department of Computer Science and Engineering. It is a privilege to submit this thesis under their guidance, constructive suggestions and steadfast support have been invaluable. I extend my heartfelt thanks to the esteemed faculty members of the Department of Computer Science and Engineering for their unwavering support and encouragement. Their advice and collaborative spirit have enriched my academic experience and contributed significantly to my personal and professional growth. I would also like to acknowledge the continuous support and encouragement provided by **Prof. Prateek Sharma**, Vice-Chancellor. His dedication to fostering a research-oriented environment has been a significant driving force behind my accomplishments.

Finally, with a heart full of love and longing, I offer my deepest gratitude to my family and friends for their unwavering support, even during the most challenging times. My Parents, Parent-In Laws, Husband, Brother, Sister, and Family & Friends have always been the pillars of my strength, and their constant support has helped me reach this stage in life.

This acknowledgment is a humble testament to the collective efforts and support of all these individuals, whose contributions have been pivotal to the successful completion of my doctoral research.

Pavi Saraswat

2K21/PHDCO/05

ABSTRACT

The rapid digital transformation and extensive adoption of social media platforms have revolutionized the advancements in mental health monitoring and depression detection. Social media platforms like Twitter, Facebook, Instagram, Reddit, etc., has firmly established itself as an indispensable part of life for the majority of the population nowadays. The constant presence of users on these platforms provides a rich user generated content that can be leveraged to monitor the mental health in comparison to traditional clinical settings. Utilizing advancements in artificial intelligence, natural language processing and computer vision, researchers and clinical mental health experts can detect early signs of depression by analysing text, audio, video, image and emoticons content generated by users on social media platforms.

This thesis presents a novel deep learning-based frameworks for depression detection using the user generated English multimodal social media content. Initially a dataset was created for the same as one of the main constraints was the non-availability of the multimodal dataset for depression detection. Additionally, a framework was presented that is a combination of bidirectional encoder representations from transformers and convolutional neural networks. This model was designed to detect the depressive posts using the created dataset, also a new model is presented to detect the severity of the post using publicly available dataset.

The research further introduces a deep learning-based framework for depression detection for hindi and hinglish (code-mixed) dataset. To overcome the challenge of regional diversity a hindi and hinglish language-based dataset was created from openly available social media platforms. This model was a combination of bidirectional encoder representations from transformers and particle swarm optimization based optimized convolutional neural network.

A comprehensive experimental evaluation is conducted to assess the performance and scalability of the proposed frameworks. The Comparative analyses with existing methodologies are carried out to demonstrate improvements in detection accuracy, efficiency, and overall system robustness. It provides a practical foundation for mental health monitoring and secure healthcare applications in real-world deployments.

List of Publications

Journal Publication

[1] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CNN approach for depression detection on social media using multimodal data. *The Computer Journal*, *67*(7), 2453-2472. https://doi.org/10.1093/comjnl/bxae018, **[Paper Published – January 2024] [SCIE Indexed]**

[2] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CPSO model for multi-class depression detection using pure hindi and hinglish multimodal data on social media. *Computers and Electrical Engineering*, *120*, 109786., https://doi.org/10.1016/j.compeleceng.2024.109786 **[Paper Published - October 2024] [SCIE Indexed]**

Conference Publication

[1] Saraswat, P., & Beniwal, R. (2024, July). BERT-based RNN for Effective Detection of Depression with Severity Levels from Text Data. In *2024 IEEE Symposium on Wireless Technology & Applications (ISWTA)* (pp. 52-56). IEEE. **[Paper Published – July 2024] (Scopus-Indexed)**

[2] Saraswat, P., & Beniwal, R. BERT Based RNN for Effective Detection of Depression with Severity Levels from Text Data, IEEE Symposium on Wireless Technology and Applications. IEEE, 2024. **[Accepted and Presented] (Scopus-Indexed)**

Communicated Articles

[1] Saraswat, P., & Beniwal, R. An Automated Hybrid Model for Depression Detection based on Vocal Features using Social Media Data , Arabian Journal of Science and Engineering, Springer. **(Under Review)**

[2] Saraswat, P., & Beniwal, R. Computer Assisted Optimised Hybrid Deep Ensemble Approach for Depression Analysis based on Facial Video Data, The Journal of Super Computing, Springer. **(Under Review)**

[3] Saraswat, P., & Beniwal, R. DViTRoM: A Transformer-Based Multimodal Framework for Depression Detection Using Text and Visual Cues, Sādhanā, Springer **(With Editor)**

Table of Content

List of Abbreviations

WHO	World Health Organization
DALYs	Disability Adjusted Life Years
AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
TL	Transfer Learning
DT	Decision Trees
DL	Deep Learning
CNN	Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transformers
KNN	k Nearest Neighbor
RF	Random forest
BiGRU	Bidirectional Rated Recurrent Units
EEG	Electroencephalography
MSMTC	Medical Social Media Text Classification
RNN	Recurrent Neural Network
LR	Linear Regression
DCNN	Deep Convolutional Neural Network
HI	Happiness Index
OCR	Optical Character Recognition
RoBERTa	Robustly Optimized BERT Pretraining Approach
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
NB	Naïve Bayes
MNB	Multimodal Naïve Bayes
TAN	Transductive Adversarial Network
BF	Bayes Factor
FRT	Friedman's Rank Test
MNB	Multimodal Naïve Bayes

GRU	Bidirectional Rated Recurrent Units
DNN	Deep Neural Network
ARNN	Autoregressive Neural Network
FC	Fully Connected
ANN	Artificial Neural Network
NB	Naïve Bayes
MNB	Multimodal Naïve Bayes
TF-IDF	Term Frequency - Inverse Document Frequency
SSP	Subsequent Sentence Prediction
GB	Gradient Boosting
ME	Maximum Entropy

List of Table

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Overview

In the current era of remarkable technological developments, the identification of an abnormal mental health condition is of the highest concern and attracting the interest of researchers worldwide. In our society, the importance of mental health is comparatively less than physical health. It is still not discussed openly because of social stigma or lack of acceptance due to understanding and awareness. Because of the ignorance of symptoms of mental disorders, it goes undetected for longer, further intensifying and leading to mental distress, discomfort and irritation of the patient. People suffering from mental disorders have to suffer from continuous stress, negative thoughts, and anxiety which affect their quality of life immensely [1]. Further, unidentified mental disorders can further lead to the self-harming approach of the patient or the worst-case scenario can lead to suicide, because of which it becomes more critical to detect it as early as possible and provide the medical help as required. Among all, depression is the most common mental illness worldwide and is generally confused with mood fluctuations or transitory emotional responses to primary life challenges [2]. Depression is considered a most perilous illness that can adversely affect someone's emotional state as well as their physical well-being. Therefore, the aim of this research is focused on efficient and automated diagnosis of such individuals, so that correct medical interventions and therapies can be adopted at the earliest, which in a way would be a step towards ensuring a good quality of life for such individuals.

The current clinical evaluation of depression and other mental disorders heavily relies on conducting clinical interviews and standardized tools for screening

such as CES-D (Center for Epidemiological Studies Depression) questionnaire and DASS (Depression Anxiety Stress Scales) questionnaire, which have variations like DASS-42 and DASS-21 for measuring depression scale through a list of questionnaires [3-4]. These measures are simple and their scientific validity has been proved by many research studies too; but they often lead to poor diagnosis due to their high dependency on just subjective and descriptive analysis. These methods are time-consuming, and people had the peer pressure to perform better rather than being realistic. Due to stigma and incorrect interpretation, patients may underreport or overstate their symptoms, and physicians may assign differing scores depending on their understanding or experience. Their precision is further limited by language and differences in culture as different populations may express and view feelings of discomfort in extremely distinct ways. Employing these scales exclusively carries the risk of faulty diagnosis, inappropriate therapy, and a superficial understanding of the patient's situation [5]. This is where the importance of social media platforms comes into the picture, as people feel free to express their genuine feelings on these platforms, which can be a great environment to collect data in abundance for the analysis of the mental state of users.

Social media has increased in significance as a platform where individuals can share their sentiments, emotions, and feelings in a number of ways. It offers an extensive amount of data that could be useful to detect signs of depression. The social sites like Facebook, YouTube, Instagram, Reddit, and Twitter are gaining importance in terms of interaction, exchange of knowledge, and thought sharing across a spectrum of sectors [6]. Further, various Machine learning (ML) and Deep learning (DL) methods have been extensively utilized to predict mental health such as depression by analyzing the text content of social media posts. Generally, the content posted on social media or the internet is highly unstructured in nature as it is published by the user, who is not a trained professional. Therefore, ML and DL based techniques are much needed to classify this kind of data. Employing such approaches has been demonstrated to significantly improve the ability to detect depression in many studies. A study by Kour [7] revealed that there is no effective way to categorized depressed and non-depressed persons, making it quite challenging. Thus, the use of ML and DL can provide best

solution to tackle depression related diagnosis.

In whole, the difficulty of accurately and consistently identifying depression emphasizes the necessity of quantifiable and objective techniques to facilitate early detection using social media platforms and automated diagnostic tools. Instead of replacing existing clinical methods, the development of such analysis aims to overcome the limitations of subjective rating scales by offering more reliable and accurate detection, which would eventually improve patient outcomes and life expectancy.

1.2 Depression: Overview and Brief

The mental health of a person is an integral part of health; it's the foundation for psychological, emotional and social well-being, which then decides a person's behavior throughout the lifetime. According to World Health Organization (WHO) report almost a billion people, which includes 14% of the World's youth were facing the problem of mental disorder in the year 2019. Depression and anxiety have existed for a more extended period of time, but by the end of the pandemic's 1st year it was increased by 25% [8]. WHO estimated that in India, the mental health problems is 2443 disability-adjusted life years (DALYs) per 100 00 population and the suicide rate is 21.1, which is leading to substantial economic loss as well [9].

Depression is the most common mental illness worldwide and is generally confused with mood fluctuations or transitory emotional responses to primary life challenges. It can be the reason for the poor performance of the person at work or family and if not detected and treated within time, it can lead to the worst possible outcome like suicide. Suicide is the 4th leading cause of death in 15–29-year-olds; wherein over 700 000 suicides occur every year. And because of this, WHO's Mental Health Gap Action Programme (mhGAP) have covered depression as its priority [10]. To analyze and detect depression at early stages, it is necessary to consider the key concepts related to depression as provided in this Section.

1.2.1 Types of Depression

Depression is a broad and diverse mental health condition that includes multiple subtypes, each with specific characteristics, causes, and methods of treatment. An outline of the primary and most common types of depression that can be seen in individuals are provided below [11-12]:

a) **Major Depressive Disorder (MDD):** MDD, also known as clinical depression, is characterized by the symptoms of frequent low mood, loss of interest in tasks, and alterations in appetite, difficulty falling asleep, drowsiness and a sense of feeling inadequate. In order to diagnose such type of depression, at least five symptoms must last for two weeks or longer. MDD is the major cause of disability which adversely impacts the functioning of people in performing daily activities.

b) **Persistent Depressive Disorder (PDD):** PDD, previously termed dysthymia, is marked by persistent and continuous depressive symptoms that remain for at least a two-year period. The symptoms, such as low self-worth, trouble paying attention, and dismay, are often more persistent but less severe compared to the signs of MDD. People with PDD might experience episodes of deep depression, a condition named "double depression."

c) **Bipolar Depression:** Alternating episodes of discouragement and hypomania, or mania, are the crucial indicators of bipolar depression type. While manic instances are marked by high mood, greater activity, and sometimes dangerous actions, depressive episodes are identical to MDD.

d) **Seasonal Affective Disorder:** A subtype of depression identified as seasonal affective disorder is more prevalent during the colder months when there is little exposure to sunlight. Insufficient energy, fatigue, eating in excess, and gain of weight comprise a few of the symptoms. Conventional treatments involve medication, psychotherapy, and light therapy.

e) **Postpartum Depression:** This subtype of depression especially impacts women after giving birth to a baby. It is characterized by extreme grief, fret, and fatigue, and may render it hard for them to take care of others. Medication, support groups, and counseling are available forms of treatment for such types of depression.

f) **Premenstrual Dysphoric Disorder:** This type of depression is characterized by significant changes in mood. The menstrual cycle symptoms during the luteal stage include nervousness, depression, and irritability.

g) **Psychotic Depression:** A severe type of depression referred to as psychotic depression is defined by psychosis, involving hallucinations or illusions. These psychotic symptoms, such as guilt illusions, frequently have
a sad tone. Antidepressants and antipsychotic medications are frequently prescribed in conjunction for treatment.

h) **Atypical Depression:** Atypical depression is identified by emotional responses and symptoms including food cravings, too much sleep, and susceptibility to rejection. It usually starts in childhood and is especially common in women.

i) **Melancholic Depression:** This subtype of depression is identifiable by a loss of happiness in most actions, a lack of flexibility to enjoyable stimuli, profound depression, and major weight loss or sleep deprivation.

j) **Recurrent Brief Depression:** Brief repeated depression episodes lasting from two to thirteen days and happening frequently a year constitute a feature of this type of depression. These episodes, though short, can be powerful and encompass thoughts of committing suicide.

1.2.2 Causes of Depression

Depression is a widely growing complicated condition which is impacted by a number of factors including biological, psychological, and social. Analyzing such factors in detailed manner is necessary for Effective prevention and early treatment of

depression. A brief description of these factors is provided as under:

a) Genetic and Biological Factors: The primary biological causes of depression include immune system, neurochemical, hormonal, and hereditary factors. The possibility of developing depression is greatly affected by genetic predisposition. MDD has a genetic estimate ranging from 30% to 50%, based on studies of twins and families [13]. Unbalanced neurotransmitters also play an essential role; depression symptoms are significantly linked to lower or imbalanced levels of norepinephrine, serotonin, and dopamine [14]. Further, neuroendocrine issues specifically hyperactivity of the hypothalamic-pituitary-adrenal (HPA) axis, result in higher cortisol levels, which limit neurogenesis and give rise to mood disorders. The impact of chronic inflammation has been reinforced by recent studies, that show that higher pro-inflammatory cytokines (such IL-6 and TNF-α) influence mood and brain function and could potentially be a factor in treatment-resistant depression [15].

b) Psychological Factors: Depression contains psychological causes that involve early memories or experiences, traits of personality, and mental operations. Cognitive theory suggests that individuals who often think in a negative way including gloomy, critical of themselves, or low self-worth thoughts, are more prone to depression [16]. Personal traits also contribute; people with high anxiety are more inclined to experience depression because they are more susceptible to emotional instability and stress [17]. In addition, an adversity in early life, such as neglect or abuse that occurs through childhood, also hinders emotional and psychological growth and makes people more vulnerable to stresses, which elevates the risk of depression [18]. Understand the helplessness, or a feeling that one cannot influence what happens in life, is a further significant psychological component. It often arises from repetitive exposure to unregulated stimuli, which promotes negativity and depression [19]

c) Social Factors: A wide range of environmental and personal factors are regarded as societal causes of depression. Major life assaults like separation and divorce, joblessness, economic distress, or grief are extremely associated with bouts of depression [20]. Furthermore, it is well-known that depression risk is elevated by

social isolation and a lack of genuine social support, with isolated individuals expressing a higher incidence of feelings of depression [21]. Another significant social indicator is a low socioeconomic position; poverty and related tensions raise the likelihood of depression, primarily due to limited access to services and chronic stress [22]. Cultural influences such as judgment, stigma, and social standards around mental health may prevent people from obtaining treatments thereby making depression's psychological costs harsher [23].

1.2.3 Symptoms of Depression

Depression develops as a mix of emotional, cognitive, physical, and behavioral indicators that fluctuate in level and duration. To accurately identify a person with depression, it is critical to pay proper attention to these signs to enable early treatment and therapies.

a) Emotional Symptoms: People suffering from depression often show emotional symptoms. The primary emotional features include constant sensations of sadness, emptiness, dismay, and tearfulness. Depression is further defined by an extensive lack of interest in almost all activities, which leads to significant decline in a person's ability to engage in daily life. Individuals frequently communicate sentiments of feeling worthless and undue guilt, which aggravate psychological discomfort and social exclusion [24].

b) Cognitive Symptoms: Cognitive symptoms have an immense impact on one's ability to thrive in professional, social, and educational settings. Attention problems, uncertainty, anxious thoughts, and obsessions are symptoms of common mental illnesses. Suicide ideas and recurring thoughts of death are possible among depressed people. Because of their significance for safety and the need of treatment, such symptoms are highly crucial to investigate [16].

c) Physical Symptoms: People with depression typically define issues with their bodies, which may at times overpower emotional concerns. According to the American Psychological Association [25], these physical signs include fatigue, a lack of energy,

shifts in appetite or significant weight swings, and sleep conditions such as sleeplessness or a state of hyper.

d) Behavioral Symptoms: Depression often leads to notable behavioral changes. Affected individuals may exhibit social withdrawal, reduced engagement in previously enjoyable activities, neglect of self-care, and significant reductions in work productivity [21]. These behavioral symptoms contribute to further social isolation, which can worsen the depressive state and complicate recovery.

It is essential to keep in consideration that people of all ages and cultures exhibit signs of depression in distinct ways. Along with a rise in behavioral issues and academic challenges, depression in children often presents as frustration rather than grief. In contrast, depression in seniors occasionally appears as physical symptoms. Additionally, the public display of depressive feelings is significantly affected by cultural factors. Instead of publicly expressing feelings of unease, depression can show up more frequently in certain cultures as physical symptoms like pain or fatigue. Further, depressive symptoms may arise in a broad range of ways, from single episodes to permanent patterns.

1.3 Depression Analysis

Depression is becoming one of the social problems as the number of sufferers is increasing on a daily basis. Mental health issues come under the category of widely accepted challenges in the World, with more than 300 million people currently suffering from depression itself. Millions of individuals globally deal with depression, a widespread and debilitating mental condition. Since early detection improves outcomes of therapy and decreases the condition's social impact, accurate and timely analysis of depression is crucial. Depression analysis includes a wide range of techniques for understanding, identifying, and evaluating depressed conditions. This section provides an overview of the approaches that are being utilized and developed in recent years.

1.3.1 Traditional Clinical Approaches

Traditional clinical methods of identifying depression depend on structured assessments and standard tools that have been effectively verified in both research and clinical scenarios. These approaches mainly consist of clinician interviews and questionnaires for self-report implied to assess the presence and extent of depressive symptoms. One of the widely used clinical tools is the Structured Clinical Interview for DSM Disorders (SCID), which is viewed as the gold standard for identifying depression by systematically examining symptoms with DSM criteria [26]. The Patient Health Questionnaire-9 (PHQ-9) is another useful self-report measure as it is short and complies with DSM-5 criteria. This clinical method assesses the frequency of symptoms related to depression over the two weeks prior to the test and is comprised of nine items with values ranging from 0 to 3 [3]. Along with extensively utilized tools such as the PHQ-9 and SCID, another two well-known self-reported scales to evaluate depression are the CES-D and DASS. A 20-item self-reported measure termed the CES-D was established to assess signs of depression in people in general. The frequency of symptoms throughout the previous week is the primary concern, and replies range from "Very rarely or none of the time" to "Most or all of the time." Additionally, there are two different versions of DASS: the brief 21-item DASS-21 and the original 42-item DASS-42. Each subscale in every version comprises 14 items derived from the DASS-42 and 7 items from the DASS-21, which measure three distinct negative feelings of stress, anxiety, and depression [4].

Furthermore, a well proven measure is the Beck Depression Inventory-II (BDI-II), comprising 21 questions that analyze the psychological, emotional, and physical signs of depression. A total score ranging from 0 to 63 is achieved by rating each item on a scale of 0 to 3. Depression scores from 0–13, 14-19, 20-28, and 29-63 indicate minimal, mild, moderate, and severe depression levels [3]. In addition, through a planned interview, clinician-administered measurements, including the Hamilton Depression Rating Scale (HAM-D), provides an objective evaluation of the severity of depression. Each of the 17–24 questions on the HAM-D are assessed on a scale, and the total of the scores reveals the level of

depression severity.

Despite the effectiveness of clinical procedures in medical settings, recent literature presents several drawbacks. The subjective nature associated with self-reported evaluations is one key problem. Due to social standards, lack of knowledge, or stigma, patients may hide their signs, which could result in a false diagnosis. Unreliable diagnoses can also result from various clinician assessments that are impacted by the clinical scenario, the practitioner's capability, and prejudicial views. Biases based on cultural and demographic factors provide further difficulties. The use of many methods for diagnosis may be constrained because they were primarily developed and validated in Western, Caucasian communities. Additionally, diagnosing symptoms of depression gets more difficult by their fluctuations [3]. There are multiple ways that depression may show up, and distinct individuals will have various mixes of indications. Finally, an absence of objective indicators for depression indicates the necessity of advanced methods of diagnosis. Despite continuing investigations into genetic markers, neuroimaging, and other biological signs, there are currently no reliable tests, making diagnosis dependent on subjective evaluations. In conclusion, whereas traditional clinical approaches have provided a basis for interpreting and diagnosing depression, their limitations underline the need for improved, unbiased, and culturally relevant tools for diagnosis [27].

1.3.2 Modern Technological Approaches

Modern techniques to identify depression have been gaining popularity as due to of the limitations of traditional diagnostic methods, such as subjectivity, symptom variability, and a lack of regular monitoring. These objective and quantitative approaches employ AI, ML, wearable sensors, online social media data, and digital biomarkers to offer more flexible and reliable diagnosis of depression. With the enormous amount of available user-generated content on social-media networking platforms, depression detection using machine learning and deep learning has become the trending topic in the research industry. As multimodal data (text,

images, and video) is evolving on social networking platforms, it has become important to explore every data posted by a user to conclude whether the user is depressive or not. Further, these results can also be mapped with the situations such as what level of depression intensity leads to suicide and how help can be provided if prior detection of depression takes place [28].

For over ten years, social media has emerged as a vital element of youths' lives. To be able to accurately convey their emotional conduct through various data, many depressed individuals frequently utilize social media for support. This information can be shared in several ways, including text, images, videos, and speech, and the analysis of such data critically can save a person's life. Teenagers are mostly active on social media platforms; there is a rise in the suicide rate of youth between the ages of 15 to 29 years. Also, it is observed that one student commits suicide every hour when he indicates it through messages like ending life or can't survive on social media platforms [29]. There have been many incidents in the recent past where people have live-streamed their suicides on social media and all of them were found to suffer from depression before they took the drastic decision of ending their life [30-32].

Since COVID-19 has affected the World, people have shifted more towards online platforms to share their feelings, leading to extensive use of social media platforms like Facebook, Instagram and Twitter. These social media sites help people to express their views, moods status, and emotions and also help them to share them with friends and family. The daily lives and mental state of a person is easily reflected through these posts, which is rich content for researchers to study a person's mental state and wellness [33-35]. A survey specified that teenagers and young adults suffering from depression or depressive symptoms are switching to social media to express their views Figure 1.1 [36]. In addition to sharing their feelings and views, multiple studies show that people use social media platforms to give advice and seek help related to health problems [37-40].

To transform the process of self-report questionnaires to digital technologies, various techniques and tools are utilized by the researchers including

ML and DL that can precisely analyze social media activity. To correctly group depression at various severity levels, these algorithms can extract extensive details from complex user data [41]. Employing such approaches has been demonstrated to significantly improve the ability to detect depression in many studies. A study by Kour [8] revealed that there is no effective way to categorized depressed and non-depressed persons, making it quite challenging. Thus, the use of ML and DL can provide best solution to tackle depression related diagnosis. The exploration of the state-of-the-art in this direction revealed a number of ML and DL techniques for detecting depression utilizing different modalities and datasets.

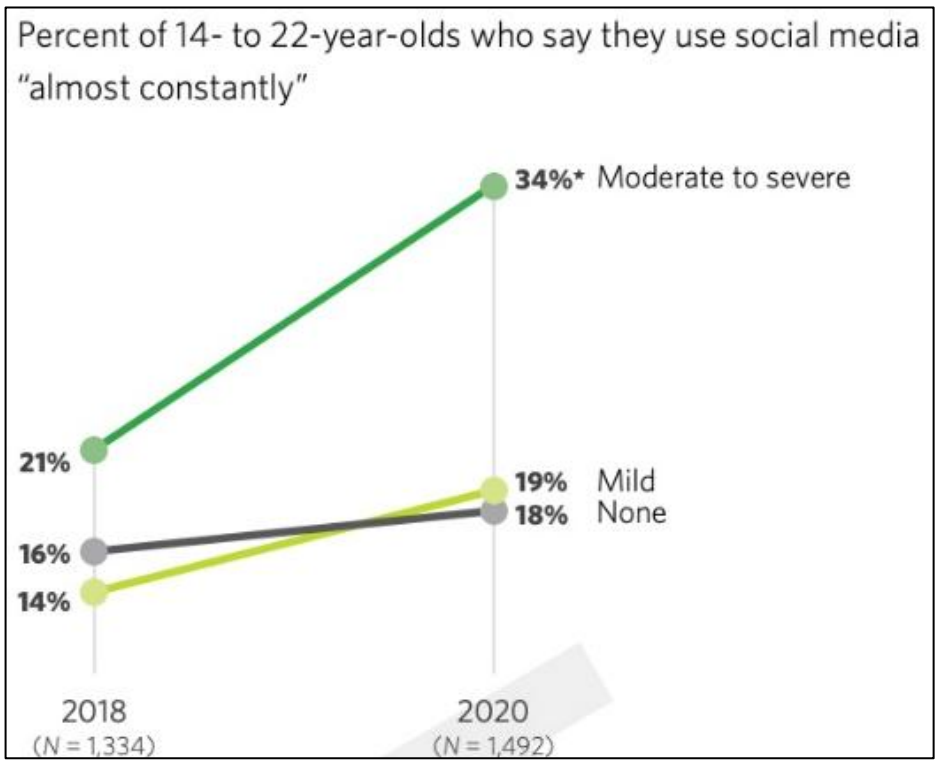


Figure 1.1: Frequency of social media use by depressive symptom levels, 2018 and 2020

Recently Machine learning (ML) and Deep learning (DL) methods have been extensively used to predict mental health by analyzing the text content of social media posts. Generally, the content posted on social media or the internet is highly unstructured in nature as it is published by the user, who is not a trained professional.

Therefore, the methods mentioned above are much needed to classify this kind of data. Deep learning has contributed a lot to several application problems like mental illness detection, stock market prediction, traffic accident prediction and many more.

Deep neural networks are more robust than typical neural networks and sometimes, they face the issue of overfitting and taking a much longer time to model the underlying data. However, it is still considered as an evolution in the area of sentiment analysis. The very first efficacious deep learning algorithm was Restricted Boltzmann Machines (RBM) which resulted in faster training as compared to other previous approaches. Then later on, Convolutional Neural Networks (CNNs) were proposed, which grabbed the popularity in image processing. It showed improved discriminative power, and also it helped to extract features along with training the data [42]. CNN generates a progressive hierarchy of abstract features through convolution, pooling, tangent squashing, rectifier and normalization as it contains some convolution stacks [43]. CNNs are generally used for the analysis of video and image patterns rather than the decoding of temporal information. Recurrent Neural Networks (RNNs) are preferred over CNNs for sequential and temporal information analysis because of their better discriminative power on these types of data. Later, Long Short-Term Memory (LSTM) was introduced in RNN to handle vanishing gradient problems, which was observed during the analysis of high dimensional time-sequential data [44].

Nowadays, the most common form of content a user generates is posts and comments on social media platforms such as Facebook, Instagram, Twitter, Reddit and many more. These posts are a combination of text, images, gifs, videos or a combination of them. These platforms have initiated a modern way of communication for people in today's World and allow a person to know about other's life and state of mind. Thus, the content on these platforms can be a rich source of data that depicts the users' feelings, emotions and mental conditions. Further analysis of this data can be used to detect depression using machine learning and natural language-based systems to decipher and comprehend users' feelings and expressions on social media platforms [45]. There are multiple articles that show the machine learning and

deep learning methods to detect depression on social media platforms, but only through text or questionnaires [46-51]. Further, for multimodal data sets only a few articles showcase the approach of sentiment analysis but for the Twitter platform only [52-53].

1.4 Overview of Depression Recognition Framework

The word "depression" is often used in connection with a mental health condition that has a substantial influence on a person's actions, mood, and overall well-being. In clinical diagnosis, behavioral diagnosis, and psychological evaluations serve as crucial tools to identify indicators of depression and patterns. The individuals who suffer from depression typically have changed sleeping routines, reduced drive, drowsiness, and memory problems. Many times, the word "depression" refers to a mental illness that has a major effect on a person's behavior, mood, and general functioning. The identification of depression symptoms and patterns in clinical diagnosis relies heavily on behavioral analysis and mental health evaluations. Depression frequently causes changes in sleep habits, decreased motivation, exhaustion, and cognitive impairment. If these symptoms are thoroughly evaluated, more effective treatment approaches may be set up, which will eventually enhance the individual's quality of life (QOL).

In other words, examining depression has significance for detecting, planning, and helping individuals affected. Depression assessment and investigation provide several benefits in the clinical domain, including:

- Aiding in the clinical decision-making procedure in order to develop efficient therapies for people with diverse emotional and psychological needs.
- Assisting in choosing the most efficient psychoactive or therapeutic strategies.
- Supporting the development of methods that mitigate the threat of emotional distress and self-harm.
- Offering quantitative information that helps track progress and facilitate medical evaluations.

While there are a variety of techniques and methods for investigating both normal and depressive people, the automated ML and DL platforms might prove a better fit for this type of research based on social media data. Each diagnostic system's framework is comprised of a number of interrelated components that combine to provide the desired result. As shown in Figure 1.2, the general architecture followed in this work to classify depressed and non-depressed individuals comprises several stages, including: dataset acquisition; extracting data, data pre-processing; feature extraction and classification; performance evaluation; and output.

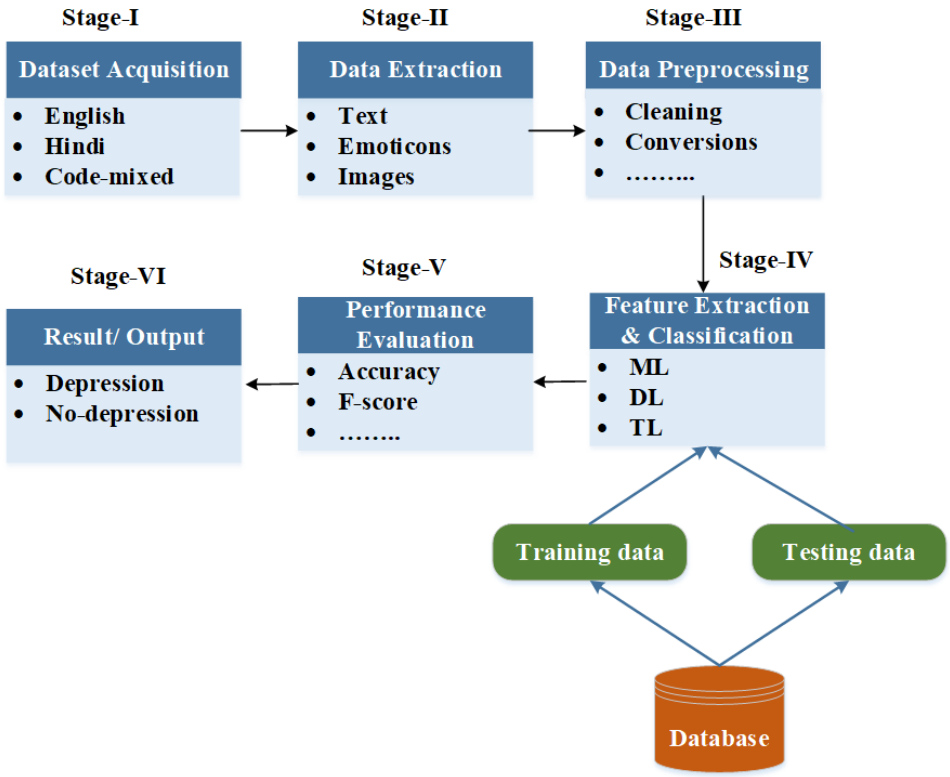


Figure 1.2: The generic framework of the diagnostic process for depression recognition

Stage I involves collection of relevant datasets consisting of depressed and non-depressed individuals from different social media platforms such as Instagram, Twitter, Facebook, etc. The data or content posted by users is in different languages such as Urdu, Hindi, Bangla, English, and many more [53-54]. Among all, this work focused and gathered English, Hindi, and code-mixed data to perform robust

multimodal depression analysis. Also, the regional languages were considered because this work focuses on the multilingual aspect of the dataset for the depression analysis. In Stage II, extraction is performed on the acquired data to extract text, emoticons, and images separately which need to be processed further as this work focuses multimodal depression analysis. So, all the different modalities are collected separately and then ensemble models are utilized in different combinations of modalities like, text +image for English, Hindi and Hinglish language. Furthermore, Audio and Video modalities were explored in English language for depression analysis. Stage III involves pre-processing all of the data and transforming it into a form that is useful and easy to comprehend when an extensive quantity of data has been acquired. URLs, stop words, spaces, hyperlinks, capital letters, and other features were among the information collected. Therefore, this work employed a variety of preprocessing techniques for analyzing text, emoticons, and images independently in order to improve the quality of the raw data. Various such techniques are used, including cleaning, eliminating stopwords, special characters, numbers, punctuations, spaces, links, redundant phrases, and converting data into a form that is both important and meaningful. After preprocessing, Stage IV which involves the extraction of features and classification of training and testing data for depressed and non-depressed posts are performed using ML, DL and Transfer learning (TL) based approaches such as Convolutional Neural Network (CNN) [42], BERT (Bidirectional Encoder Representations from Transformers) [41], RoBERTa (Robustly Optimized BERT Pretraining Approach), DistilBERT, XLNet, ResNet, VGG Net, MobileNet, etc. Furthermore, multiple ML techniques are also explored for classification purpose like SVM (Support Vector Machine), DT (Decision Tree), NB (Naïve Bayes) and KNN (K-Nearest Neighbour). These classification techniques are employed separately on both text and image data to extract relevant features in-depth and classify data successfully. Additionally, their hybrid models like BERT-CNN, BERT-CPSO (PSO Optimized CNN). BERT-SVM, BERT-KNN, BERT-DT, BERT-NB, BERT-ResNet, BERT-VGGNet, BERT-MobileNet are explored and compared for the best performance measures. Further, to evaluate the performance of the diagnostic models and perform a robust comparative analysis, various evaluation measures are adopted in Stage V considering Accuracy, F1-score, Precision, Recall, etc. Finally, Stage VI indicates the result of the proposed

diagnostic system for depression identification based on multimodal data and accordingly provides a decision regarding the presence and absence of the depression.

1.5 Research Gaps

From the literature survey, the following research findings have been identified:

1. In the past few decades, researchers have been using interview-based and questionnaire-based methods to analyze the behaviour and patterns in the mood or languages of a depressed user. These methods are lengthy in terms of processing and expensive in nature. Furthermore, it becomes tough to collect an adequate amount of data that will guarantee the robust and generalized character of the models, which predicts the presence of depression in a user.

2. The indicators or markers of depression also evolve with the progression of technologies, so after the growing nature of social media, it has attracted more users toward it and is an excellent pool for content. But there are multiple modalities (text, image, and video) of data posted on these platforms daily, and the past research has focused chiefly on textual data and visual data is comparatively less explored.

3. However, it has also been noticed that depression is a disorder that may exist in people at different levels or intensities. At the initial stage, the intensity is low and its severity increases with time. Also, the treatment is diverse for each stage. Therefore, the intensity of the depression also plays an important role.

4. One of the major issues is that there are no such frameworks or models for depression detection which work on all modalities like text, image, tags, emoticons and videos. Also, as all the social media platforms have different ways of accessing their modalities, there is no common model/framework that can work on all the trending social media platforms by considering all the modalities. Some of the main

challenges are listed below regarding this problem:

- There is no publicly available large-scale benchmark multimodal dataset for depression detection on social media platforms.

- Users' contents and behaviors on social media are unstructured and heterogeneous. It becomes tough to illustrate the users from different perspectives and capture the relation across different modalities.

- Even though users' behaviors are rich and diverse, only a few symptoms (in terms of keywords, Hashtags, emoticons, etc.) tend towards depression. This makes the depression-oriented features scarce on social media platforms and problematic to be captured.

5. As it is a well-known fact that English is the most commonly used language on social media platforms, many researchers have created a model with text modality for depression prediction on social media platforms. But the usage of local languages is not far behind as it is also apparent that India is a diverse country.

By addressing these gaps, future research can significantly enhance the effectiveness of depression detection using social media content.

1.6 Problem Statement

Accurate depression detection remains a critical challenge, with most existing studies relying solely on English text data, neglecting the rich insights offered by multimodal content and underrepresenting regional languages like Hindi. Research seldom explores the integration of text and image data, and there is a notable lack of publicly available Hindi multimodal datasets. Moreover, deep learning techniques such as transfer learning and optimization remain underutilized in this domain.

This work aims to develop a robust, multimodal deep learning classifier that leverages both text and images to categorize social media posts into Depressed

and non-depressed classes. By addressing language limitations, incorporating visual cues, and applying advanced learning strategies, the goal is to significantly enhance the accuracy and reliability of automated depression detection systems.

1.7 Research Objectives

Based on the literature review of existing state-of-the-art methods and the research gaps identified in the above sections, the following research objectives are significant and suitable for further developing a Smart and Secure Healthcare System.

RO1: To perform the systematic literature review of the multimodal deep learning models for depression detection from social media posts.

RO2: To develop a multimodal deep learning-based framework for detecting depression by affective analysis of Social Media posts from different platforms for Multimedia Dataset.

RO3: To develop a multimodal deep learning-based framework for detecting depression by affective analysis of Social Media posts for the Hindi language content.

RO4: To do a comparative result analysis of the developed models with other existing models/techniques.

1.8 Outline of Research Outcomes

The results indicate the highest utilization rates for combined modalities (37.4%), the Twitter dataset (37.2%), and deep learning (DL) techniques (41.90%), compared to other modalities (text, speech/audio, image/video), datasets (Facebook, Weibo, AVEC, Reddit, DAIC-WoZ, combined, and others), and learning approaches (machine learning and hybrid methods). Furthermore, a new multimodal dataset of Instagram and other social media platform posts, annotated with the help of clinical experts, was created and used to classify depressive vs. non-depressive posts, filling a major gap in available public datasets. The study proposes a robust hybrid approach combining BERT for text analysis and CNN for image analysis to detect depression

from Instagram and other social media platforms posts, achieving state-of-the-art accuracy. BERT achieved 97% accuracy on text data; CNN achieved 89% accuracy on image data; and the hybrid BERT-CNN model achieved 99% accuracy, outperforming traditional ML and other hybrid combinations like BERT-SVM, BERT-KNN, and BERT-DT. In addition, this thesis work proposes a robust hybrid BERT-RNN-based model to classify depression into three categories: not-depressed, moderately depressed, and severely depressed. The proposed method achieved a high classification accuracy of 95%, outperforming existing techniques and state-of-the-art models. Also, this work developed a novel Hindi and Hinglish-based multimodal depression dataset using data from social media platforms. A hybrid BTCPSO model combining BERT for text and PSO-optimized CNN (CPSO) for images is proposed, achieving 97% accuracy in depression detection. Further, BERT outperformed its variants in text analysis with 95% accuracy. PSO-enhanced CNN (CPSO) achieved 95% accuracy, outperforming basic CNN and other ML/DL methods for image data. A detailed description and analysis of all the research outcomes obtained are presented in the next chapters one by one.

1.9 Organization of Thesis

Chapter 1: Introduction

This chapter will introduce the background, problem statement, research objectives, and the significance of developing deep learning-based depression detection models using multimodal social media content.

Chapter 2: Literature Survey

This chapter will provide a comprehensive review of the existing work on the multimodal deep learning models for depression detection from social media posts, highlighting existing gaps, challenges, and recent advancements.

Chapter 3: Multimodal Deep learning-based frameworks for detecting depression using English social media content *(Aligned with RO2)*

This chapter will focus on developing and implementations of deep learning framework(s) for depression detection using self-acquired multimodal English dataset with details of techniques used in the proposed models .

Chapter 4: A Deep learning-based framework for depression detection using multimodal Social Media posts for the Hindi language content *(Aligned with RO3)*
This chapter will discuss the design and implementation of a deep learning-based framework for depression detection of social media posts on self-acquired Hindi and Hinglish code mixed dataset.

Chapter 5: Conclusion, Future Work, and Social Applications
This chapter will summarize the research's key findings, contributions, and limitations. It will also suggest directions for future work along with the social impact.

1.10 Chapter Summary

This chapter showcases the overview of the remarkable technological advancements in identification of depression with help of social media posts using ML, DL and TL based ensemble models. Furthermore, it provides an overview and brief of depression that includes, types of depression, causes of depression, symptoms of depression followed by depression analysis. Both the traditional clinical approaches and modern technological approaches are briefed. Then, an overview of depression recognition framework is presented followed by research gaps which will be addressed through a systematic literature review presented in Chapter 2. Moreover, Problem statement, research objectives, outline of research outcomes and organization of thesis are provided for a clearer understating of the thesis.

In conclusion, this chapter lays a foundation for a detailed literature review that addresses the research gaps and provides a groundwork for developing deep learning frameworks using multimodal social media content for depression detection.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

Depression is considered a most perilous illness that can adversely affect someone's emotional state as well as their physical well-being. Depression itself accounts for 10% of all disabilities worldwide associated with physical as well as mental health problems. It may seriously alter the ability of an individual to a greater extent in performing regular activities, like working, eating, sleeping, and appreciating life [1]. One of the primary drivers of suicides in the globe is depression. As per data from the World Health Organization (WHO), there are over 800,000 verified instances of depression-associated suicides every year. More than 26% of individuals stated signs of depression within a year of adhering to the COVID-19 pandemic. In addition, it is expected that depression will be ranked as the second most prevalent reason for disability globally by 2030 [54]. The most prevalent signs noticed in persons suffering from depression include high stress levels, changes in mood, fear, lack of desire, self-harm thoughts, difficulties with decision-making, forgetfulness, etc.

Social media's significance has increased as a platform where individuals can share their sentiments, emotions, and feelings in a number of ways. Also, there are multiple other aspects for using online network data such as personality detection by leveraging social psychology, mental health monitoring, behavioral pattern analysis, lifestyle and interest profiling etc. Hence, social media offers an extensive amount of data that could be useful to detect signs of depression. For over ten years, social media has emerged as a vital element of youths' lives. The social sites like Face-book, YouTube, Instagram, Reddit, and Twitter are gaining importance in terms of interaction, exchange of knowledge, and thought sharing across a spectrum of sectors [6]. To be able to accurately convey their emotional conduct through various data, many depressed individuals frequently utilize social media for support. This

information can be shared in several ways, including text, images, videos, and speech, and the analysis of such data critically can save a person's life. Clinical evaluation for depression that make use of lengthy procedures such as rating systems, surveys, and interviews may be inaccurate, leading to excess time and erroneous analysis [5]. These subjective medical procedures depend extensively on the patient's interaction capacities as well as the doctors' knowledge. To avoid social dislike, patients may consciously disguise their actual perceptions from specialists. Thus, computerized analysis is crucial for detecting depression efficiently.

To transform the process of self-report questionnaires to digital technologies, various techniques and tools are utilized by the researchers including Machine learning (ML) and Deep learning (DL) that can precisely analyze social media activity. To correctly group depression at various severity levels, these algorithms can extract extensive details from complex user data [41]. Employing such approaches has been demonstrated to significantly improve the ability to detect depression in many studies. A study by Kour [7] revealed that there is no effective way to categorized de-pressed and non-depressed persons, making it quite challenging. Thus, the use of ML and DL can provide best solution to tackle depression related diagnosis. The exploration of the state-of-the-art in this direction revealed a number of ML and DL techniques for detecting depression utilizing different modalities and datasets. To perform further investigations, it is necessary to analyze the amount of past work done on these aspects.

In today's era, the usage of social media is not limited to a certain number of persons or geography; however, people from different cultures, countries, languages, etc. use it abundantly to share their thoughts freely. As people use different languages to convey their sentiments, language is an essential factor that should be taken into account. Therefore, different languages such as Bengali, Urdu, Malay, English, etc. are preferred by former researchers to carry out their work as given in Table 2.1 to improve depression diagnosis. A novel multi-class Urdu database that includes user evaluations is provided by Khan et al. [55] for sentiment analysis. The data provided was collected from a variety of industries and comprised 9312 hand-

compiled reviews divided into three categories: favorable, negative, and neutral by specialists. One of the research project's targets was to develop reference solutions employing rule-based, ML, and DL methodologies. Also, BERT was adjusted in multiple languages for the evaluation of sentiment in Urdu language. Their concluded that the DL was surpassed by the proposed BERT model and a total F1 score of 81.49% was attained with rule-driven and ML algorithms. Therefore, this chapter primarily focused on modalities, datasets, learning techniques and languages employed by the researchers for depression detection. A total of 47 research articles are selected to be considered in the chapter for review using different databases such as Google Scholar, PubMed, Springer, Elsevier, etc. the main contributions of this paper are summarized in the points below:

1) Exploration of different type of modalities and their utilization in detection of depression.
2) Systematic analysis of various datasets available and experimented by researchers to detect depression.
3) Exploration of learning platforms preferred by the past studies to effectively detect depression.
4) Analysis of the languages of the datasets preferred by the researchers in the previous studies.
5) Understanding the drawbacks and future scope of the existing studies.

2.2 Research Questions

To carry out a systematic survey a well-defined research questions was formulated in this phase. The established research questions create essential foundations for identification and evaluation of research. The detailed research questions are mentioned below:

1) What are the modalities being widely utilized for depression detection?
2) Which datasets have been used in past years to analyze depression effectively?
3) Which learning techniques have been frequently practiced by researchers to detect depression?

4) What languages have been utilized in past years for analysis of depression detection?
5) What are the research gaps identified in this survey article and future perspectives that need to be covered for more efficient depression detection?

2.3 Search Strategy

Artificial Intelligence (AI) tools are being studied by researchers from recent years to detect depression using data on social media. However, a very less number of systematic surveys on the detection of depression are observed. Therefore, in this survey article, the main focus is directed towards the detection or analysis of 'De-pression' considering different modalities, datasets, learning techniques and languages. The data from reputed journals and conferences is gathered after proper refining the articles through inclusion and exclusion criterions. To search the relevant articles, a systematic process is followed as shown in Figure 2.1, in accordance to "Preferred Re-porting Items for Systematic Process and Meta-Analyses" (PRISMA) guidelines [56]. The PRISMA process comprises of different phases starting from articles identification to articles selection which resulted in a total of 47 related and relevant articles used in this review. Initially, distinct databases are searched including Google Scholar, PubMed, Springer, Elsevier, Web of Science, the Multidisciplinary Digital Publishing Institute (MDPI) Library, the Institute of Electrical and Electronics Engineers (IEEE) Library, etc. for detailed literature from (2013-2024). The use of various keywords in the search box such as 'depression', 'suicide', 'stress', 'anxiety', 'de-pression detection', 'depression analysis', 'depression diagnosis', 'depression modality', 'depressed text', depressed images', depressed speech', 'depression datasets', 'depression databases', 'ML for depression', 'DL for depression', 'AI for depression detection', etc. provided with 344 articles.

Further, after limiting and refining the search string, removing duplicates, irrelevant, and mismatched articles, 254 papers are obtained. Another round of screening (case reports, newsletters, mag-azine articles, non-human, summary

documents, etc.) is performed which yielded 74 articles. Then, full-text articles were assessed for the eligibility and a proper exclusion criterion is set. The articles: whose objective was not to detect or analyze depression, have incomplete data or information, and have not applied ML and DL are removed (n=12) and n=59 is found to be eligible for inclusion. Finally, out of 59, a total of 47 most relevant and related articles are selected to be included in this review. Therefore, it is believed that the time given to review huge data of articles during this chapter will be helpful for further updated research on depression detection.

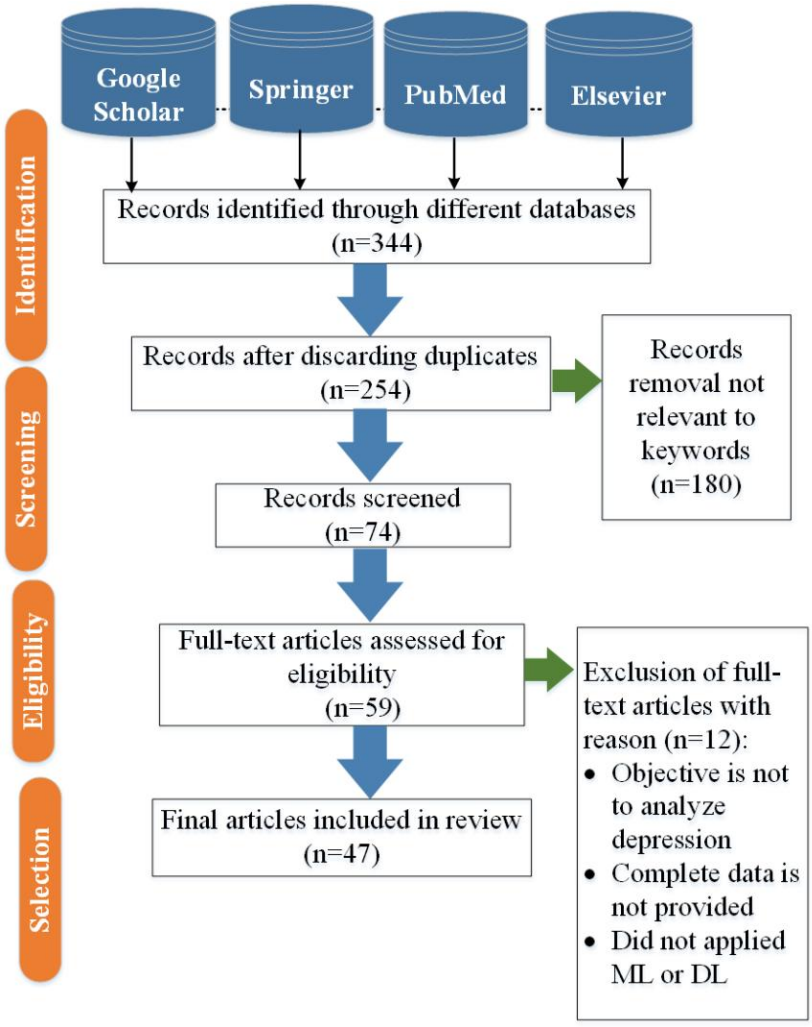


Figure 2.1: PRISMA flow-chart depicting the articles search strategy

2.4 Methodology

This section provides a comprehensive literature survey of techniques and sources utilized for effective depression detection as tabulated in Table 2.1. The purpose of this chapter is fivefold. Firstly, the state-of-the-art indicating the usage of depression detection modalities is conducted. Secondly, the datasets utilized in the past studies related to depression are explored. Thirdly, the usage of different ML and DL techniques employed in the literature is scrutinized to detect depression accurately. Fourthly, the languages used in the past studies in regard to depression detection is discovered. Fifthly, based on the literature performed, gaps are identified and future suggestions are made.

2.4.1 Modalities for Depression Detection

In recent years, researchers utilized a number of modalities including audio, text, and images/videos to perform robust depression detection. Some research works combined multiple modalities to improve the accuracy of depression detection. The main research works highlighting the usage of these modalities is provided in this sub-section.

2.4.1.1 Based on Audio Data.

The Transformer feature extraction interpreter and model on Deep Neural Network (TIF-DNN) using audio data was developed by Biradar et al. [57]. Their study employed the IndicNLP and English-to-Hindi modules for transcription and BERT for extracting features. The proposed model's accuracy of 73% was the best when compared to the baseline methods to recognize hate speech. How-ever, the proposed translator-based DL system can efficiently predict depression but regional languages and a more robust translator system need to be included. Du et al. [58] incorporated the linguistic changes of depressed individuals into a variety of common audio factors to generate an audio based on a chained structure. Experimentation on the Distress Analysis Interview Corpus/Wizard-of-Oz (DAIC-WOZ) dataset, time-

domain features were retrieved employing Recurrent Neural Network (RNN), yielding accuracy and F1 scores of 77.1% and 74.6%, respectively. The results proved the superiority of the model in diagnosing audio-based depression but have very small sample size.

In an investigation by Chlasta et al. [59], an approach to recognize speech depression employing residual Convolutional Neural Network (CNNs) and audio modality was presented. The findings of the network analysis of the five layouts revealed the efficacy of "ResNet-34 and ResNet-50" in classification. A successful approach for audio spectrograms analysis of disturbed individuals appears in the results; yet, more testing on balanced datasets is needed. A supervised learning method for detecting social depression and anxiety in audio data was presented by Salekin et al. [60]. Neural Network 2 vector (NN2Vec) was utilized that identify and exploit inherent connections between speakers' vocal states and their symptoms. Also, a framework is proposed that integrates NN2Vec features with the Bidirectional Long Short-Term Memory and multiple instances learning (BLSTM-MIL) classifier, to detect speakers with a high degree of social anxiety as well as depression with F-1 scores of 90.1% and 85.44%, respectively. Future research will enquire at the viability and security of the used strategy.

2.4.1.2 Based on Text Data

Utilizing DL models on text data, Amanat et al. [61] built a reliable method for early depression detection. CNN, Support Vector Machine (SVM), Naïve Bayes (NB), Decision Trees (DT), and RNN-LSTM models are fed using data in textual form taken from the Kaggle repository. The research results revealed that, in contrast with different approaches, the proposed RNN-LSTM obtained the greatest accuracy of 99.6%, however, it has the drawback of having a smaller sample size. For the purpose to assess depression, Gupta et al. [62] investigated the efficiency of techniques for classification on both balanced and unbalanced data. Twitter and Kaggle are utilized to gather text data. In comparison with other ML methods, the chosen LSTM model showed superior accuracy, precision, and recall of 83%, 84%, and 75%, respectively.

For more efficient analysis, the work may be expanded by integrating ML and DL platforms. With a 79% accuracy rate, Arora et al. [63] employed SVM and Naïve Multinomial Bayes (MNB) on text data to study depressed users. The result therefore highlighted the critical role of ML in detecting depression. The testing just using text data and neglecting the degree of depression are among the disadvantages of the study. To detect anxiety indications, Kumar et al. [64] created a list of words related to anxiety. To train the model, three models were used: MNB, Random Forest (RF), and gradient boosting (GB). Further, an ensemble classifier was employed for majority voting. The proposed structure succeeded in classifying text data with an accuracy of 85.09%. In the future, depression analysis techniques that are verified on a different user population must be investigated.

Singh et al. [65] presented a study that makes use of Natural Language Processing strategies to detect emotions in mixed data, including text composed in Hindi and English. The methods' efficiency was evaluated across multiple datasets and attained a satisfactory accuracy rate of 76.6%. The used cluster-based strategy worked well to process code-mixed data, but it was unable to identify emotions in long words. A technique to identify hate speech in Devnagri Hinglish texts was given by Chopra et al. [66]. They developed a classifier based on a Tabnet model, which surpassed other methods' accuracy, reaching 90%. Therefore, the proposed model has shown effectiveness at detecting depression in text; however, the segmentation procedure requires improvement. An idea was offered by Chekima et al. [67] to tackle some of the issues that Malay online text frequently raises. The suggested approach significantly improved the accuracy by retaining 79.28% of the data in comparison to the baseline's 51.38%, although it required a significant amount of human interaction. NN, RF, 1-DCNN, and SVM were employed by Mustafa et al. [68] to generate a multi-classifier that successfully classifies texts on depression. The proposed DL model 1-DCNN achieved the best efficiency of all the classifiers, with an accuracy of 91%. A combination of features could improve diagnostic performance in the future.

Sentiment analysis regarding the Bangla language was the main focus of Bhowmik et al. [69]. They created an exclusive domain-related categorical vocabulary

dictionary of data and a new rule-based procedure called Bangla Text Sentiment Score (BTSC) was used. They employed the appropriate BTSC scores and the term frequencies-inverse on the two datasets, then two matrices were created which were then evaluated using supervised ML classifiers. Their results revealed that the support vector machine (SVM) received the best recognition accuracy of 82.21% exhibiting the usefulness of the BTSC approach in Bangla sentiment assessment.

2.4.1.3 Based on Image/Video Data

By combining two three-dimensional CNNs, Melo et al. [70] tried to increase the efficacy of the diagnostic process. The spatiotemporal interactions are generated in both the local and global face areas using the AVEC dataset's image and video data. The method is advantageous but involves a lot of data to train the model. Employing video data for depression level analysis, Jazaery et al. [71] developed a model approximation for the Beck Depression Inventory grade. The approach detects the spatiotemporal attributes of the face region using an RNN-C3D network. The mean-absolute-error (MAE=7.37) and root-mean-square error (RMSE=9.28) were obtained also the issues related to human behavior need to be considered.

The presented model thus produced improved results but more A two-way deep network approach, the Inception-ResNet-V2, was presented by Uddin et al. [72] utilizing video data to get dynamic features and spatial information. The proposed approach to detect depression performed better than other methods. In the future, model improvements can be made to significantly enhance RMSE and MAE rates. CNN was applied by Chao et al. [73] and trained on the image and video-based AVEC 2014 dataset, producing RMSE and MAE of 9.98 and 7.91, respectively. The proposed method's limitation was its extremely small image size, yet it was capable to accurately estimating the depression score. Kamalesh et al. [74] offered to employ ML approaches to perform personality analysis via social media platforms. A novel system comprising of a Binary Partitioning Transformer, Frequency, and Inverse Gravity Moment is presented. On the Facebook, Instagram, and Twitter datasets, the proposed model's accuracies were highest at 78.34%, 79.67%, and 86.84%, respectively. The

presented model outperformed all other models showing its efficiency in depression analysis but used only image data.

2.4.1.4 Based on Combined Data

Research by Wu et al. [75] used a combination of da-ta, such as textual, behavioral, and environmental factors, to diagnose depression early. With accuracy, recall, and F1 scores of 83.3%, 71.4%, and 76.9%, respectively, they presented an automated and effective DL approach to analyze depression.

In order to assess the task of identifying emotions in difficult scenarios, Ruz et al. [76] combined text data, emoticons, and tags. On Twitter datasets, they employed a range of ML, including NB, RF, SVM, Transductive Adversarial Network (TAN), and Bayes factor (BF) on TAN. With an accuracy of 81.2%, the findings showed that SVM worked best; although, more training samples have to be taken into account. An efficient multi-tasking depression analysis model based on double Bidirectional Rated Recurrent Units (BiGRU) was offered by Han et al. [77]. They obtained SentiDrugs' mixed data, which included text, emoticons, and tags. The results obtained showed that their suggested model had the highest accuracy, at 78.6%. For an in-depth review, video data should also be taken into consideration. A DL-based system that can classify emotions as either extremist or non-extremist via analysis of text, tags, and emoticons was put forward by Ahmad et al. [78]. They evaluated the suggested hybrid CNN and LSTM against various ML and DL techniques, including KNN, RF, SVM, NB, CNN, and LSTM. With an accuracy of 92.06%, the proposed model delivered the most accurate results; however, contextually aware factors were neglected. Using text, emoticons, and tags in Facebook, Katchapakirin et al. [79] designed a Natural Language Processing (NLP) structure for depression diagnosis in Thai. They employed LSTM and attained an 85% accuracy rate. The study concluded that DL-based techniques are crucial tools to analyze depression levels from the behavior of Facebook users.

The CNN model achieved an accuracy of 88.4% when Lin et al. [80]

employed it on Twitter data that had both text and image features. Deshpande et al. [81] used ML models on text, and emoticons to analyze symptoms of depression. The findings showed the successful classification of de-pression using MNB and SVM with 83% and 79% accuracy, respectively. An encoder-decoder prediction framework was offered by Das et al. [82] for categorizing Bengali posts from Facebook pages that consist of emoticons and text. An attention system, LSTM, and decoders based on gated recurrent units (GRUs) were employed for creating the model. At last, the attention-based decoder approach scored the highest accuracy of 77% when compared to the other three encoder-decoder net-works. Cummins et al.'s study [83] identified spatial-temporal and gradient attributes from voice, image, and video data. In order to generate histograms, these features were processed additional using the bag-of-word approach. Support Vector Regression (SVR) was finally used for training and testing, and an RMSE of 10.65 was achieved. One of the drawbacks is that the technique used for audio failed to work well under test settings. In research by Jan et al. [84], the regression model is used to analyze depression based on auditory and visual data. The results demonstrated that depression had been well classified, with an RMSE of 10.32 and an MAE of 8.16. For better outcomes, additional dynamic and acoustic information must be retrieved in the future.

Table 2.1: Literature Survey of Depression Modalities, Datasets, and Learning Techniques for English language content

Author (Year)	Modality	Dataset	Learning Techniques	Language	Performance	Remarks
Anshul et al. [85] (2024)	Text, Image and URL	Tweets-Scraped dataset	MFEL (Multimodal Feature based Ensemble learning) (proposed), LR, XBG, NN	English	Accuracy: LR = 88.3% XGB = 89.7% NN = 86.4% MFEL = 91.7%	• Leveraging a combination of textual, visual, and user-specific features proposed model outperforms others • In future, other modalities can be included.
Sadhegi et al. [87] (2024)	Text, Semantic features, Facial data	E-DIAC, PHQ-8	Bi-LSTM, DepRoberta	English	MAE= 2.85 RMSE – 4.02	• The best results were achieved by enhancing text data with speech quality assessment,

						• Future work can be extended using hybrid approaches and visual/image data.
Du et al. [58] (2023)	Audio	DAIC-WoZ	DL: RNN	English	Accuracy=77.1% F1-score= 74.6%	• The results proved the superiority of the model in diagnosing audio-based depression • Very small sample size
Han et al. [87] (2023)	Audio	DAIC-WoZ and AVEC 2019	DL: RNN	English	Accuracy=78 %	• The proposed model significantly improved the accuracy of depression detection • Only audio data is experimented
Chopra et al. [66] (2023)	Text	Twitter	ML: SVM, DT, RF, XGBoost DL: DNN, LSTM, Tabnet (Proposed)	Code Mixed Data	Highest accuracy with Tabnet =90%	• Tabnet based classifier model shown efficacy in detection depression from code-mixed data • Dimensionality reduction methods need to be practiced • Segmentation needs to be improved
Nadeem et al. [88] (2022)	Text	Twitter	ML: LR, SVM DL: SSCL, CNN, GRU, LSTM, BiLSTM	English	Best accuracy with SSCL model= 97.4%	• Data automation and feature extraction by the proposed study improved the task of depression detection • Depression severity need to be focused
Kamalesh et al. [74] (2022)	Image	Twitter, Facebook and Instagram	ML: BPT, TF-IGM	English	Accuracy on: Facebook = 78.34% Twitter = 79.67% Instagram = 86.84%	• The proposed model outperformed all other models showing its efficiency in depression analysis • Use of only single modality

Belinda et al. [89] (2022)	Text	Twitter	ML: Multinomial NB	Code Mixed Data	Accuracy= 96.15%	• The proposed system based on Hindi-English language can be utilized globally in reducing depression rate • The work can be further improved by incorporating a bot for interaction
Amanat et al. [61] (2022)	Text	Twitter (Kaggle)	ML: SVM, NB, DT DL: CNN, RNN-LSTM (proposed),	English	Best Accuracy with RNN-LSTM =99.6%	• The highest accuracy is achieved by the proposed technique for depression detection • Hybrid DL models should be utilized for future analysis • Less sample size
Poświata et al. [90] (2022)	Text	Reddit	DL: RoBERTa, BERT, XLNet	English	Best accuracy with Ensemble of RoBERTalarge and DepRoBERTa = 65.8%	• The proposed work provided a wining solution for detection depression signs on social media • Model need to be trained on larger text corpus
Gupta et al. [62] (2022)	Text	Twitter (Kaggle)	ML: DT, KNN, SVM, LR DL: LSTM	English	Highest results with LSTM: Accuracy =83% Precision =84% Recall =75%	• DL based LSTM has shown highest results in detecting depression • The work can be extended by hybridizing ML and DL platforms for more effective analysis
Zogan et al. [51] (2022)	Text	Twitter	MDHAN	English	F1-score= 93.4%	• The fusion of DL and multi-aspect attributes shown effective way to diagnose depression • The study only considered Twitter data
Khan et al. [55]/ (2022)	Text (comments and	Urdu Corpus for Sentiment	fastText (Bi-LSTM, Bi-GRU, CNN-	Urdu	Highest accuracy with the	• The study created a dataset for Urdu sentiment analysis.

	reviews)	Analysis (UCSA-21)	1D, LSTM, GRU, CNN-1D+MP, CNN-1D+ATT, LSTM+MP, LSTM+ATT) and Proposed (BERT)		proposed BERT = 81.49%	• The article deployed multiple ML and DL algorithms for multi-class classification and in future GPT2, GPT3 etc., can be explored.
Bhowmik et al. [69]/ (2021)	Text	Cricket and Restaurant dataset	LR, KNN, RF, SVM on UniGram model	Bangla	Accuracy of Restaurant data: SVM= 77.91% LR = 70.41% KNN = 69.41% RF = 66.14% Accuracy of Cricket data: SVM= 78.69% LR = 72.81% KNN = 63.25% RF = 65.26%	• The research proposed Lexical data dictionary for Bangla language (LDD). • Less number of samples are included which affects the model's performance.
Singh et al. [65] (2021)	Text	Twitter	ML: NB, SVM DL: LSTM	Code mixed data	Best accuracy with ensemble =76.6%	• The presented cluster-based method can be useful for handling code-mixed data • Less number of samples are included • Model failed in detecting emotions in long sentences
Biradar et al. [57] (2021)	Speech	Twitter	DL: TIF-DNN	Code-Mixed data	Accuracy= 73%	• The proposed translator-based DL system can efficiently predict depression • Regional languages and more robust translator system need to be included

Dai et al. [91] (2021)	Audio, Semantic features and Video	DAIC-WoZ	DCNN and DNN	English	F1-score= 96%	• The proposed approach provided excellent accuracy for depression detection • Feature selection phase was time-consuming • Few samples were included from database
Das et al. [82] (2020)	Text and emoticons	Facebook	DL:CNN+LSTM, CNN + GRU, CNN+ARNN (Proposed),	Bangla	Highest accuracy with Proposed model= 77%	• Attention-based decoders have the high capability in analyzing depression efficiently • The existing data need to be increased • Detecting type of speech written can improve the diagnostic accuracy
Han et al. [77] (2020)	Text and Emoticons	SentiDrugs	DL: PM-DBiGRU (proposed) and others	English	Highest accuracy with the proposed model = 78.6%	• The proposed approach can enhance e dug level aspect-reviews for sentiment analysis • Video data is not considered for more robust analysis
Vazquez et al. [92] (2020)	Audio	AVEC 2016	DL: Ensemble of 1D-CNN	English	Accuracy: 68%-74%	• The fusion of networks offered a promising way for automatic depression detection • Bagging and boosting techniques need to be explored more in future • The combination of modalities can be focused to improve accuracy rate
Wang et al. [93] (2020)	Text and Images	Weibo	ML: SVM, RF, LR, NB, GBDT, AB, FusionNet (Proposed)	Chinese	Highest F1-score with proposed =0.9772	• The proposed model proved to be ideal solution when handling multiclass depression problem • The size of dataset

						needs to be improved • Further analysis of user behaviour is required
Uddin et al. [72] (2020)	Image/ Video	AVEC 2013 and AVEC 2014	DL: Inception-ResNet-V2 model	English	On AVEC 2013: RMSE= 8.93 MAE= 7.04 On AVEC 2014: RMSE= 8.78 MAE= 6.86	• The proposed technique outperformed other methods for depression detection • The RMSE and MAE rates can be further improved by making alterations in model
Lin et al. [94] (2020)	Audio+ Text	DAIC-WoZ and AViD-Corpus	DL: CNN, Bi-LSTM	English	Accuracy: 89%-90%	• The fusion of networks offered a promising way for automatic depression detection • Visual data need to be included for in-depth analysis
Wang et al. [95] (2020)	Tags, Text, and Emoticons	Weibo	ML: SVM, NB, LR DL: BERT (proposed), CNN, LSTM	Chinese	Highest Accuracy with Proposed =75.6%	• The proposed BERT can be efficiently used in sentiment analysis. • Only a single platform is considered for data analysis • Disease Severity is not focused
Ruz et al. [76] (2020)	Text, Emoticons	Two twitter datasets	ML: NB, SVM, RF, TAN, BF	English	Highest Accuracy with SVM =81.2%	• The properties of SVM seemed to be effective in depression analysis • More number of training samples need to be considered
Mustafa et al. [68] (2020)	Text	Twitter	ML: NN, RF, SVM DL: 1DCNN	English	Highest accuracy with 1DCNN =91%	• The diagnostic performance can be improved using a combination of features to detect depression • More number of modalities should

						be focused
Chekima et al. [67] (2020)	Text	Facebook and Twitter	ML: Proposed system, baseline, NB, SVM, ME	Malay	Best accuracy with Proposed model= 79.28%	• The proposed system has shown improved accuracy than baseline methods for Malay based depression detection • Limitation of frequent lexicon adjustment • High human intervention
Wu et al. [75] (2020)	Text, Behavior and Environmental traits	Facebook and CES-D	DL: D3-HDS, Word2vec + LSTM	English	Precision =83.3% Recall =71.4% F1-score =76.9%	• The proposed DL technique has shown effective results in early identification of mental illness • More evaluation metrics should be considered
Uddin et al. [96] (2019)	Text	Twitter	DL: LSTM	Bangla	Accuracy =86.3%	• The findings of the study will be helpful for doctors to analyze the behaviour of depressed users from their social activates • Need experimentation on large datasets
Arora et al. [63] (2019)	Text	Twitter	ML: SVM, MNB	English	Accuracy = 79%	• The outcome demonstrated the significance of ML in depression detection • Experimentation is performed on text data only • Severity of depression is not considered
Ahmad et al. [78] (2019)	Text, Tags. Emoticons	Twitter	ML: KNN, RF, NB, SVM DL: CNN, LSTM,	English	Highest Accuracy with proposed method =	• The highest accuracy achieved by the proposed CNN+LSTM model reveals its

			CNN+ LSTM (proposed)		92.06%	significance in depression analysis • Contextual-aware parameters were ignored • Non-consideration of multiple modalities
Chlasta et al. [59] (2019)	Audio	DAIC-WoZ and AVEC 2017	DL: Residual CNNs	English	Accuracy= 77%	• Results revealed an efficient method for audio spectrograms of disturbed persons • More tests need to be conducted on balance datasets • Combination of models should be experimented in future
Kumar et al. [64] (2019)	Text	Twitter	ML: LR, SVM DL: SSCL, CNN, GRU, LSTM, BiLSTM	English	Highest accuracy with ensemble =85.09%	• The presented model can correctly predict the anxious depression • More modalities need to be explored for depression analysis • Model need to be validated on distinct user base
Melo et al. [70] (2019)	Image/ Video	AVEC 2013 and AVEC 2014	DL: C3D+3D-GAP network	English	On AVEC 2013: RMSE= 8.26 MAE= 6.40 On AVEC 2014: RMSE= 8.31 MAE= 6.59	• The proposed model yielded best results than other methods depicting its efficacy in depression detection • A large amount of training data is required for the proposed model • Study needs to be explored considering more healthcare concerns
Salekin et al. [60] (2018)	Audio	AVEC 2013 and AVEC 2014	DL: BLSTM-MIL+NN2Vec ML:	English	For social anxiety: Accuracy= 90% F1-score=	• The combined approach has achieved best results than all baseline models

			NN2Vec		90.1% For state anxiety: Accuracy= 90.2% F1-score= 93.4%	• Future work includes examining feasibility and safety of the employed approach
Jazaery et al. [71] (2018)	Image/ Video	AVEC 2013 and AVEC 2014	DL: RNN-C3D	English	On AVEC 2013: RMSE= 9.28 MAE= 7.37 On AVEC 2014: RMSE= 9.20 MAE= 7.22	• The fusion of RNN and C3D can produce improved results for analysis of depression levels • More issues related to human behavior need to be considered • Study needs to be explored considering more healthcare concerns
Islam et al. [97] (2018)	Text	Facebook	ML: DT, KNN, SVM Ensemble learning	English	Accuracy: KNN = 60% Ensemble = 64% SVM = 71% DT = 71.2%	• ML approaches have shown significance in depression detection with less error • Results need to be cross verified by taking a greater number of data
Chen et al. [98] (2018)	Text, Emoticon, and Text	Twitter	ML: SVM, DL: LSTM, CNN, CNN+LSTM, RNN+CNN+ LSTM (proposed)	English	Accuracy (proposed) = 78.42%	• The designed multi-class model attained highest accuracy than existing techniques. • More accuracy can be attained by adjusting the hyperparameters • Only twitter data is focused
Katchapaki rin et al. [79] (2018)	Text, Emoticon, Tags	Facebook	DL: LSTM	English	Accuracy = 85%	• DL based technique seemed to be an important tool to analyze depression levels from behavior of Facebook users • Only Facebook data is considered

Deshpande et al. [81] (2017)	Text and emoticons	Twitter	ML: MNB, SVM	English	Accuracy: SVM= 83% MNB= 79%	• The applied ML based techniques seemed to be effective in depression analysis • Lack of tests on multimodal data • Less number of samples
Reece et al. [99] (2017)	Image features	CES-D and Instagram	ML: RF, Bayesian logistic regression	English	Recall = 67.7% F1 score = 64% Precision = 64%	• The proposed ML model shown significant results in depression diagnosis. • More information should be gathered for detailed analysis • Lack of use of multimodal data
Chao et al. [73] (2015)	Image/ Video	AVEC-2014	DL: LSTM-RNN, CNN	English	RMSE= 9.98 MAE= 7.91	• The proposed approach correctly predicted depression score using DL models on multimodal data features • Very small size of image • Reduced spatial information
Lin et al. [80] (2014)	Text, behavioral, and image features	Weibo and Twitter	ML; SVM, RF, NB DL: DNN (proposed)	Chinese	Accuracy: NB = 68.1% SVM = 75.6% RF = 76.7% DNN = 78.5%	• The proposed DNN model shown efficacy in detecting early stress in depressive subjects • Depression levels need to be focused
Jan et al. [84] (2014)	Audio+ Image/ Video	AVEC 2014	ML: Regression model	English	RMSE= 10.32 MAE= 8.16	• The presented study is cable of analyzing depression scales based on visual and audio data • More dynamic and audio features need to be extracted for improved results
Cummins et al. [83] (2013)	Audio+ Image/ Video	AVEC 2013	ML: Support Vector Regression (SVR)	English	RMSE= 10.65	• Multimodal depression analysis seems to be more accurate

						• Feature level fusion shown the promising results for both development and test sets. • The employed approach for audio didn't perform efficiently in test conditions
Almaev et al. [100] (2013)	Image/ Video	MMI and Cohn-Kanade dataset	ML: Support Vector Machine (SVM)	English	2AFC score=0.98 (highest on AU6)	• In comparison to static, dynamic features are seemed to be stronger • The LGBP-TOP approach has shown best results for correct features extraction • Levels of depression should be focused in future
Wang et al. [101] (2013)	Text, Emoticons and Images	Sina Micro-blog	ML: Bayesian rules	Chinese	Precision= 80%	• The proposed model provided a robust way to online monitor health of the users • More data need to be included in experimentation

2.4.2 Datasets for Depression Detection

To perform depression detection efficiently, researchers have utilized a number of datasets and resources comprising of different types of data. An overview indicating utilization of such datasets and data sources in previous studies is provided in this sub-section.

2.4.2.1 Audio/Visual Emotion Challenge (AVEC) dataset

To detect depression utilizing audio and visual data, researchers made use of a well-known dataset named as Audio/Visual Emotion Challenge (AVEC) dataset.

There are different versions of AVEC dataset including AVEC-2013, AVEC 2014, AVEC 2016, AVEC 2017, AVEC 2018, and AVEC 2019 which are publicly available. Among all, AVEC 2013 and AVEC 2014 are seemed to be experimented by the past research works to detect depression. For the AVEC 2013 depression dataset, 150 video clips were taken from 82 subjects in a human-machine interaction task via a camera (30 frames per second) and microphone. All of those participating in the study range in age from 18 to 63, having an average age of 31.5 and a standard deviation of 12.3 years. These clips possess a size of 640 × 480 pixels and the dataset is composed of three sets: training, development, and test. Each section includes 50 videos, each of which has a label indicating the degree of depression. The AVEC 2014 depression dataset is a modified version of the AVEC 2013 dataset, which comprises 150 video clips from the freeform and Northwind tasks. The subjects had to reply to a variety of questions in the "Freeform" task and read aloud portion from a story in the "Northwind" activity.

Melo et al. [70] experimented on AVEC 2013 and AVEC 2014, to demonstrate the usefulness of their depression detection method. Jazaery et al. [71] employed the video data from the AVEC 1013 and AVEC 2014 datasets to generate a model that can determine the degree of depression. An RNN-C3D network is implemented to detect characteristics in space and time after the face images are taken out of the video. Cummins et al.'s study [83] utilized audio, picture, and video data from the AVEC 2014 dataset to determine space, time, and histogram features. Upon training and testing with SVR, an RMSE of 10.65 was achieved. A study by Jan et al. [84] applied a regression model to evaluate depression based on data from the ACEC 2014 dataset, and the authors obtained an RMSE of 10.32 and an MAE of 8.16.

2.4.2.2 Facebook dataset

A study was put forward by Islam et al. [97] to evaluate depression using Facebook data which is accessible online. The evaluation of four supervised ML classifiers and ensemble model found that DT achieved the best accuracy, at 71.2%,

showing the algorithms' efficacy in depression analysis. Further data collection is required to cross-verify the results. Employing Facebook text data, emoticons, and tags, In order to diagnose depression in Thai, Katchapakirin et al. [79] created a NLP framework. Applying LSTM, they attained an accuracy rate of 85% and revealed the efficacy of DL-based methods to estimate the severity of depression according to Facebook users' activity. Das et al. [82] developed an encoder-decoder prediction system for Facebook post classification. When contrasted with the other three simi-lar networks, the attention-based decoder approach showed the greatest accuracy of 77%.

2.4.2.3 Weibo dataset

Utilizing text and images from the Weibo dataset, Wang et al. [93] presented a study to address the issue of depression analysis. Fusion Net is proposed for the purpose which is evaluated against other ML techniques. The proposed model proved to be an ideal solution by achieving the highest F1-score of 0.97 but the size of the dataset needs to be improved. Wang et al. [95] suggested utilizing the BERT model on Weibo to evaluate sentiment. The proposed BERT model was compared to SVM, NB, LR, CNN, and LSTM. Results indicated an ideal accuracy of 75.6% using BERT. The research could be extended in the future by employing more possible social media channels.

2.4.2.4 DAIC-WoZ dataset

A novel two-stage feature selection approach was proposed by Dai et al. [91] and applied to the high-dimensional DAIC-WOZ dataset, containing audio, video, and semantic data. Evaluation was additionally performed on the feature categories. With an F1-score of 0.96, a Precision of 1.00, and a Recall of 0.92 on the development set, the model obtained the best results in depression classification; still, feature selection requires an adequate amount of time. In their experiments on the DAIC-WOZ dataset, Du et al. [58] employed RNN to obtain time-domain features. The network

achieved F1 scores of 74.6% and accuracy of 77.1%, respectively. Despite having a relatively small sample size, the results showed the model's efficiency to detect audio-based depression.

2.4.2.5 Reddit dataset

Employing pre-trained RoBERTa models on Reddit data, Poświata et al. [90] conducted depression signs identification. An ensemble of RoBERTalarge and DepRoBERTa reached the highest accuracy of 65.8% amongst all RoBERTa and XLNet versions. In future, model need to be trained on larger text corpus.

2.4.2.6 Twitter dataset

Mustafa et al. [68] gathered information on depression by employing the Twitter platform to collect tweets. With an accuracy of 91%, the proposed multi-classifier, 1-DCNN, achieved the best results, proving its success in depression classification. A DL-based model was put forward by Chen et al. [102] to be used with Twitter data for multi-class depression categorization. When compared to the other models using DL, the proposed approach, RNN-CNN-LSTM, has the highest accuracy of 78.42%. SVM and MNB were used to twitter data by Arora et al. [63] to examine opinions and grief from their posts. With a 79% accuracy rate, their study revealed the critical role ML algorithms play in depression analysis. Singh et al. [65] presented a study that employs Natural Language Processing techniques to detect emotions in Twitter-based data. The assessment of their proposed approach on several datasets produced adequate accuracy results. Belinda et al. [89] suggested identifying people who may be sad based on their Twitter posts.

Term frequency-inverse document frequency (TF-IDF) and MNB techniques were used to classify users as either normal or depressed which showed an accuracy of 96.15%. DNN was employed by Biradar et al. [57] to detect hate speech. The pro-posed design offered the best results once it was translated from English to

Hindi via libraries. To be able to identify depression in 31,000 messages on Twitter, Nadeem et al. [88] developed a hybrid Sequence Semantic Context Learning (SSCL) system that utilizes a self-attention mechanism. For binary labelled data, the suggested SSCL yielded the highest accuracy of 97.4%. Chopra et al. [66] presented a strategy for detecting hate speech in Devnagri Hinglish texts gathered from Twitter. The created Tabnet model achieved 90% accuracy, outperforming the accuracy of prior approaches. For the purpose of identifying depression, Gupta et al. [65] evaluated the success rate of several methods of classification for both balanced and unbalanced data. Text data is obtained for experiments using the Kaggle repository's Twitter data. The accuracy of the proposed LSTM model is seemed to be greater than that of ML methods like DT, KNN, SVM, and LR.

2.4.2.7 Combined

Wu et al. [75] combined data from Facebook and The Centre for Epidemiological Studies-Depression (CES-D) to carry out a study for early detection of depression. Their DL approach, D3-HDS, achieved a satisfactory accuracy of 83.3%. Using a set of images obtained from Instagram and CES-D, Reece et al. [99] lever-aged ML algorithms to detect depression. With an F1-score of 64%, their approach highlighted the importance of ML. Chekima et al. [67] suggested an approach for dealing with some of the issues with Malay online content. By retaining 79.28% of the accuracy, the approach suggested, when applied to both Facebook and Twitter data, substantially improved it. In a study by Lin et al. [94], combined audio and video data from the DAIC-WoZ and AViD datasets were utilized. The results indicate that the suggested Bi-LSTM provides the highest accuracy for detecting depression, at 90%. In a study by Chlasta et al. [59], depression in audio was determined by combining data from the DAIC-WOZ and AVEC datasets. The effectiveness and success of residual convolutional neural networks (CNNs) in accurately classifying data has been proven by the considered study.

2.4.2.8 Others

In addition to the above mentioned, some other datasets have been utilized by the researchers to study depression. Text, emoticons, and tags from SentiDrugs were utilized by Han et al. [77] for effective depression analysis model based on BiGRU. The model they proposed showed the highest accuracy, at 78.6%. In an effort to ease automatic depression analysis employing Sina Micro-blogs that contain text, photos, and emoticons, Wang et al. [101] presented a method for data mining. 80% precision was attained by using Bayesian rules, offering a reliable method of online user health monitoring.

2.4.3 Learning Techniques for Depression Detection

To Recent years have shown the significant increase in the utilization of learning techniques including ML and DL to enable automated detection of depression. These techniques have the capability to deal with complex data in different forms and generate the desired outcome. Overview research studies focusing ML and DL for depression analysis is provided in this sub-section.

2.4.3.1 DL Methods

Zogan et al. [51] put forward a Multi-Aspect Depression Detection with Hierarchical Attention Network MDHAN to automatically identify individuals with depression on social media. Of the 4,208 users who participated in the research, 51.30 % experienced depression and 48.69 % did not. With an 89% score on F1, the combination of DL and multi-aspect features showed an efficient method of detecting depression. The work must be evaluated on other datasets in the future. With the goal to develop audio-based structure, Du et al. [58] linked the speech variations of those with depression into several common audio parts. Obtaining time-domain features by employing a DL-based method, i.e. (RNN), on the DAIC-WOZ dataset, an accuracy

rate of 77.1% is achieved. So, the results confirmed the DL model's ability to detect audio-based depression. Vazquez et al. [92] presented an automated DL-based classification system utilizing speech recordings of depression users. With a single input, four hidden, and one output layer, the system implemented 1D-CNN with n=50 for ensemble averaging. The proposed DL approach worked better than the other ML and DL techniques. The combination of modalities can be focused to improve accuracy rate.

Uddin et al. [72] developed the Inception-ResNet-V2, a two-way deep network approach that employs video data to extract attributes. The proposed method outscored other methods to detect depression. RMSE and MAE levels can be raised even more in the future via model modifications. Uddin et al. [96] utilized LSTM-based depression evaluation on the social media data. The highest accuracy rates of 78.90% with LSTM size = 128; 81.1% with batch size 25 and 10 epochs; and 86.3% with LSTM 5 layers were reached after 5,000 text tweets from Twitter were analyzed. Various DL models including RoBERTa, BERT, and XLNet are employed by Poświata et al. [90] on Reddit data to perform automated detection of depression. The results revealed the significance of the ensemble formed with RoBERTalarge and DepRoBERTa by gaining highest accuracy of 65.8%. In a study by Chlasta et al. [59], a well-known DL model, CNN, was employed to detect depression in audio data. The most promising results for diagnosing depression came from experiments on the DAIC-WOZ and AVEC datasets. A DL technique was employed in a study by Wu et al. [75] to detect depression in its early stages utilizing information from text, behavior, and environmental variables. The experiment's accuracy metric findings, which came out at 83.3%, demonstrated how well the suggested model worked to analyze depression.

Dai et al. [91] applied CNN and DCNN to evaluate the DAIC-WOZ dataset, containing audio, video, and semantic data, in a context- aware fashion. In each predicted scenario, sparse subsets were chosen using the recommended method. The ideal result on the development set is attained with a highest depression categorization score of 0.96. The research's evaluation of depression severity found

higher than that of the best reference model (RF).

2.4.3.2 ML Methods

Utilizing an image set collected from Instagram, Reece et al. [99] employed ML algorithms to detect depressive signs. RF and Bayesian LR are adopted for categorization and evaluation of strength. With a precision, recall, and F1-score of 64%, 67.7%, and 64%, their approach excelled all other approaches. Deshpande et al. [81] employed ML models on Twitter data to evaluate depression signs. To achieve this purpose, they employed MNB and SVM, and the results confirmed that these models were effective for detecting depression with the highest accuracy of 83%. The concept of data mining was employed by Wang et al. [101] to uncover individuals in distress on social media networks. They make use of human-generated rules and words in their process. Finally, the implication of Bayesian rules based on ML, achieved a precision of around 80%.

Kamalesh et al. [74] proposed a study to use ML approaches for assessing personality traits using images from social media sites. A Transformer, Frequency, and Inverse Gravity Moment formed the key elements of the proposed novel system. The proposed model's practicality in diagnostic tasks is shown by its greatest accuracy of 86.84% on the Instagram dataset. To achieve unmatched levels of real-time for detecting accuracy, Almaev et al. [100] offered the novel dynamic appearance descriptor Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP), which integrates Gabor filtering with spatial and dynamic texture evaluation. Experiments based on the MMI Facial Expression and Cohn Kanade databases show that LGBP-TOP performs superior to both of its static versions.

2.4.3.3 Both ML and DL Methods

To determine the efficacy of an employed model for depression detection, some researchers evaluated ML and DL techniques against each-other. Atlast, past studies concluded the best performing technique (either ML or DL) based on the

evaluation metric. The effectiveness of many ML and DL models, such as CNN, SVM, NB, DT, and RNN-LSTM, on text data was evaluated by Amanat et al. [61]. The study indicated that, when compared to other methods, the DL-based RNN-LSTM model achieved the best accuracy, at 99.6%. DL and ML-based models, including BERT, SVM, NB, LR, CNN, and LSTM, were employed in research by Wang et al. [95] on data containing text, tags, and emoticons. The BERT model possessed an accuracy rate of 75.6%, demonstrating the success of DL in depression prediction. To classify sentiments into two categories: extremist or non-extremist, Ahmad et al. [78] contrasted the proposed CNN and LSTM with other ML techniques, including KNN, RF, SVM, NB, CNN, and LSTM. Compared to ML models, their suggested DL model provided the best results, with an accuracy of 92.06%. Utilizing data from Twitter, Chen et al. [29] assessed the DL based proposed model, RNN-CNN-LSTM, contrary to SVM, CNN, LSTM, and CNN-LSTM. Their approach yielded the best accuracy of all, 78.42%, showing the clear advantage of the DL model combination over other approaches.

Nadeem et al. [88] developed a DL-based system, SSCL, which employs the process of self-attention to detect depression on Twitter. The proposed DL model obtained the best accuracy of 97.4% for data with binary labels. The usefulness of ML and DL approaches for the classification of depression data was investigated by Gupta et al. [62]. The accuracy of the DL-based LSTM model i.e. 83% used proved to be the highest compared to other ML techniques.

2.5 Major Findings from Literature Survey

This section provides the results of the survey performed in this chapter with the help of graphs. The analysis is made according to the different articles from 2013 to 2023. Four set of experiments have been performed to address the considered research questions as under:

2.5.1 What are the modalities being widely utilized for depression detection?

First experiment is designed to identify different modalities that are most widely utilized in previous years. The findings obtained by exploring the literature are shown in Table 2.2 and pie chart in Figure 2.2. This data shows that there are three main modalities preferred by the past studies to detect depression i.e. text, speech/audio, and image/video. Some researchers have combined these modalities to improve the overall performance of the model for depression diagnosis. Each modality is utilized in different percentage. Of all the modalities employed, the combination is seemed to be considered the most i.e. 38.29% followed by text data (34.04%), image/video data (14.89%), and speech/audio data (12.76%) respectively. The usage of combined modalities offers more realistic and robust diagnosis in comparison to a single modality. Therefore, it is utilized in higher %age and has large capability in depression prediction.

Table 2.2: Percentage utilization of depression modalities

Modality	Usage (in %)
Text	34.04%
Speech/Audio	12.76%
Image/Video	14.89%
Combined	38.29%

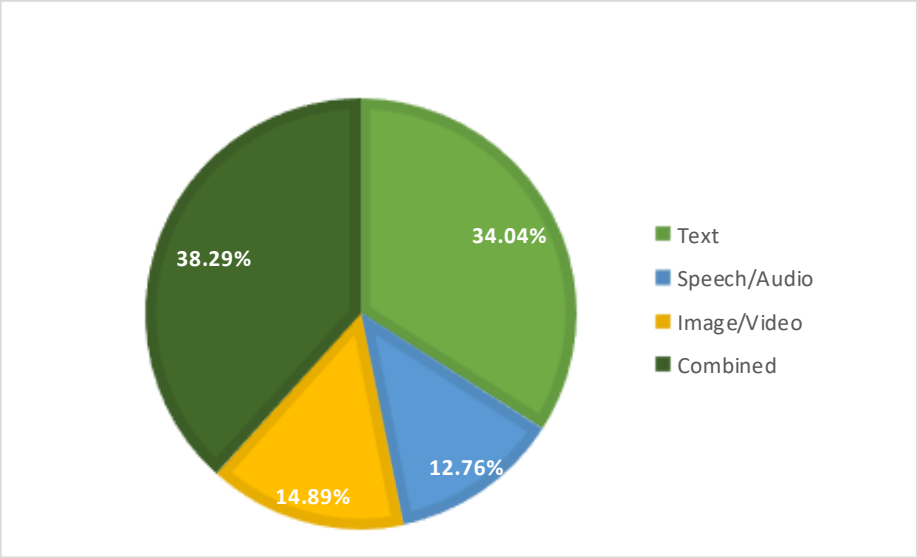


Figure 2.2: Graph showing percentage utilization of depression modalities

2.5.2 Which datasets have been used in past years to analyze depression effectively?

The second experiment is intended to determine different datasets available and utilized by researchers in past years. Table 2.3 presents the data obtained and the graphical analysis of results is provided with the help of pie chart in Figure 2.3. The findings depict a variety of datasets including Facebook, Twitter, Weibo, AVEC, DAIC-WoZ, Reddit, others, and combined that are adopted to perform depression related experimentation. So, all the datasets and data sources were from online social networking sites. Most of the emphasis i.e. 38.29% is given to Twitter data. Large availability, up-to-date information, real-time updates and easy accessibility makes the Twitter to be utilized in high %age by researchers to analyze depression. Large Twitter datasets can provide high depression accuracy when training the model. After twitter, most of the research interest is directed towards using combination of data from different datasets i.e. 19.14%. Then, the different versions of AVEC datasets are seemed to be high (17.02%). 6.38% of research focused on using Facebook and other (8.51%) datasets including SentiDrugs, Sino micro-blogs, Cohn Kanade, and MMI facial Expressions database each, followed by DAIC-WoZ (4.25%), Weibo (4.25%), and Reddit (2.12%) data sources to study depression.

Table 2.3: Percentage utilization of depression datasets

Dataset	Usage (in %)
Facebook	6.38%
Weibo	4.25%
Twitter	38.29%
Reddit	2.12%
AVEC	17.02%
DAIC-WoZ	4.25%
Combined	19.14%
Others	8.51%

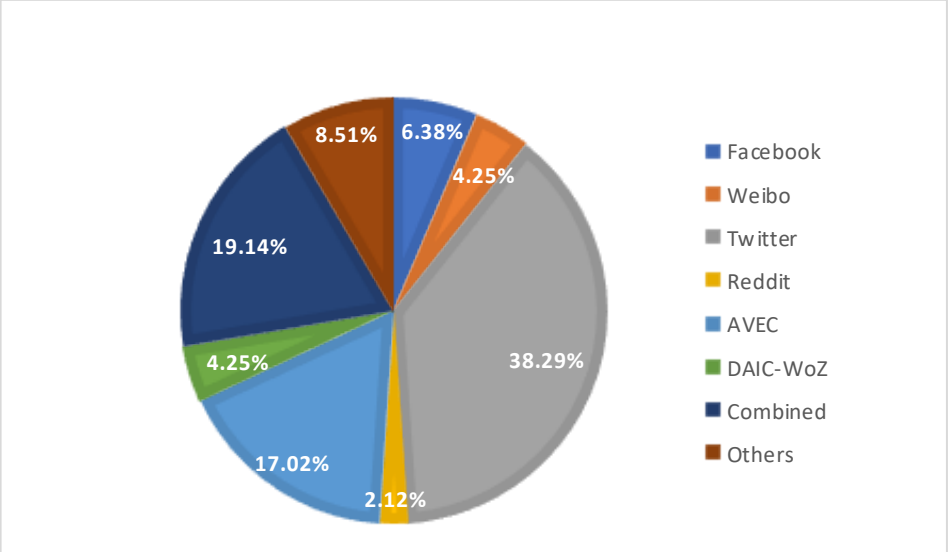


Figure 2.3: Graph showing percentage utilization of depression datasets

2.5.3 Which learning techniques have been frequently practiced by researchers to detect depression?

The third experiment comprised of identifying the learning techniques practiced by researchers to detect depression. Table 2.4 and Figure 2.4 provide the results of such experiment in the form of pie chart. The outcome shows the utilization of ML and DL in depression detection. The chart in Figure 2.4 shows that researchers have applied DL models more frequently in comparison to ML techniques. About

40.42% of the studies used DL approaches including CNN, RNN, LSTM, etc. ML for depression detection including SVM, LR, RF, NN, etc. gained about 31.91%. Some of the research works i.e. 27.65% evaluated both ML and DL against each-other to determine the best one. Thus, realistic method offered by DL techniques and their remarkable benefits including requirement of less human intervention, capability to deal with large and complex data, automated extraction of features from raw data, and adjustable parameters, make it popular among researchers to be widely used in depression detection.

Table 2.4: Percentage utilization of learning techniques for depression

Learning Technique	Usage (in %)
ML	31.91%
DL	40.42%
Both ML & DL	27.65%

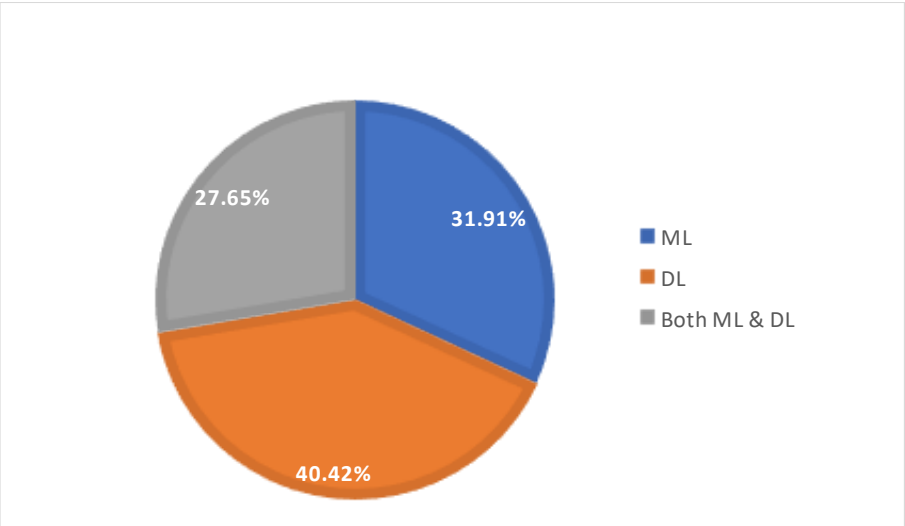


Figure 2.4: Graph showing percentage utilization of learning techniques for depression detection

2.5.4 What languages have been utilized in past years for analysis of depression detection?

The fourth experiment comprised of identifying the languages practiced by researchers to detect depression. Table 2.5 and Figure 2.5 provide the results of such experiment in the form of pie chart. The findings depict a variety of languages including English, Code-Mixed data, Urdu, Bangla, Chinese and Malay are used to perform depression related experimentation. Most of the emphasis i.e. 74.46% is given to English language which the globally accepted language. Because, it is the world-wide utilized language due to which it is utilized in high %age by researchers to analyze depression. After English, most of the research interest (8.51%) is directed towards Code-mixed data like Hinglish, which is a combination of both Hindi and English and Chinese language. Then, the Bangla language is explored (6.38%). 2.12% of research focused on using Urdu and Malay language to study sentiment analysis and depression.

Table 2.5: Percentage utilization of languages used for depression detection

Language	Usage (in %)
English	74.46%
Code-Mixed Data	8.51%
Urdu	2.12%
Bangla	6.38%
Chinese	8.51%
Malay	2.12%

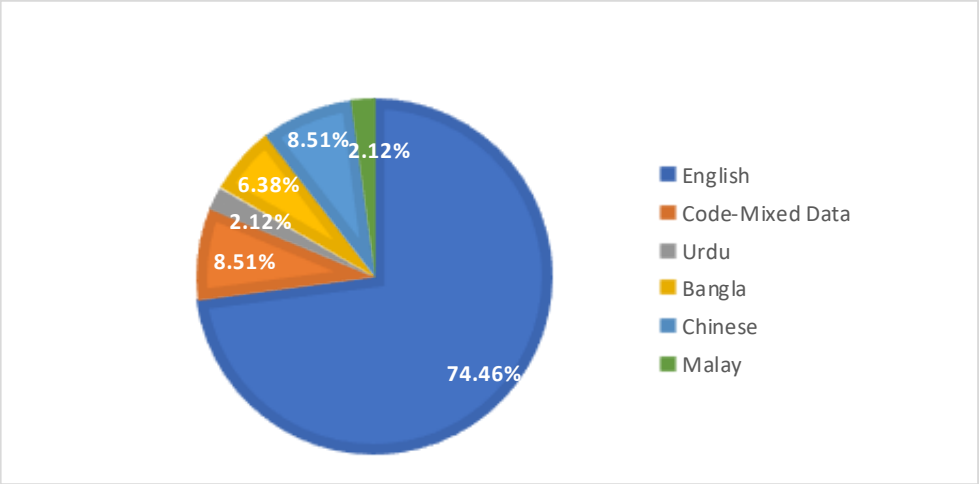


Figure 2.5: Graph showing percentage utilization of languages for depression detection

2.5.5 What are the research gaps identified in this survey article and future perspectives that need to be covered for more efficient depression detection?

The fifth experiment aimed to focus the research gaps in previous studies and provides a solution to tackle them. The comprehensive exploration of literature revealed certain gaps that need to be focused in future. An overview of such gaps and future perspectives to overcome them are presented as under:

1. Less consideration of severity levels: The detailed evaluation of a disease can't be performed until it is analyzed at all the severity levels. It has been observed that a few studies focused on evaluation of depression at different severity levels i.e. early, moderate and severe. Therefore, future work should be taken into concern regarding the examination of depression at different severity to understand the alterations in symptoms according to the stage of the patient.

2. Need of a robust dataset: To evaluate the depression efficiently, it is necessary to have a proper dataset. As already mentioned, there are number of datasets identified in literature consisting of different types of data. Still, there are certain issues to be taken care of in future. First, many of the datasets utilized in past studies are not publicly available, contain a single modality and their accessibility requires a very

lengthy process. Therefore, there is crucial need for creation of a robust dataset that include all data such as speech, video, images, text, and emoticons. Second, Twitter is seemed to be widely preferred by researchers in literature, so, data from other data sources including Instagram, YouTube, WeChat, Telegram, etc. also need to be analyzed. Third, in many studies, the drawback of small sample size is observed. As the number of samples directly impacts the accuracy of a system, therefore, in future, sample size should be increased to improve depression detection accuracy.

3. Feature Space Reduction: Very few studies in literature seemed to utilize feature space reduction techniques. Also, the chapter analyzed no to less use of a robust optimization techniques which can improve performance of the model. So, future work should be performed by using a single robust approach optimized with swarm intelligence to select optimal features and reducing dimensionality.

4. Hybrid Modelling: The incorporation of ML and DL models seemed to have high capability in detection of depression with high accuracy. In previous research works, the fusion of robust ML and DL models to each-other or with an optimized technique has not been practiced. Therefore, the amalgamation of different combination of models should be performed in future to improve the performance of the diagnostic system.

5. Need of Multilingual datasets: As it is a well-known fact that English is the most commonly used language on social media platforms, many researchers have created a model with text modality for depression prediction on social media platforms. But the usage of other languages is not far behind. Therefore, it is important to explore other regional languages too.

2.6 Performance Evaluation Metrics

This section defines the metrics used to evaluate the effectiveness and efficiency of the proposed Depression Detection Models. It includes standard

performance measures used to evaluate ML and DL models.

To build an effective ML or DL model, it is necessary to evaluate the performance of the model. The quality of the model can be analyzed using various performance evaluation metrics, which tell how well the model performs on the given data. Based on the results obtained, the hyperparameters of the model can be adjusted to enhance its performance. In this research, metrics such as accuracy, F1-score, recall, and precision are utilized to measure the performance and robustness of the proposed models with other techniques. Four main parameters, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), are used in the computation of these metrics having different meanings. TP correctly indicates the presence of a state and TN correctly indicates the absence of a state. Similarly, FP denotes the wrong indication of the presence of a state and TN is the wrong indication of absence of a state. Mathematically, the evaluation metrics can be represented as shown in equation 2.1, equation 2.2, equation 2.3 and equation 2.4 below [103].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

$$Sensitivity/Recall = \frac{TP}{TP+FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.3}$$

$$F1-score = \frac{2}{\frac{1}{recall}+\frac{1}{precision}} \tag{2.4}$$

Accuracy is the ratio of the number of accurate predictions to the total number of predictions. Recall or Sensitivity represents the number of TP outputted by the model. Precision indicates the ability of the model to identify a specific class. F1-score is the harmonic mean of precision and recall. The values of all these metrics lie between 0 and 1.

2.7 Chapter Summary

This survey was performed with the intent to highlight and explore the existing work related to depression detection. Depression being a serious mental issue needs to be identified in its early stages. To perform an extensive analysis, a total of 47 research articles are considered following the PRISMA regulations. The focus of this chapter is fivefold. Initially, the modalities being widely utilized in past works for depression detection are comprehensively analyzed. Secondly, the datasets which are available and used by researchers are studied. Thirdly, the learning techniques which have been utilized in recent years are discovered. Fourthly, the languages which have been practiced in current years are explored. Lastly, the research gaps are identified that need to be covered in future to enable early depression detection. Each of the aspect is scrutinized in detail to determine the extent of work already done. The results obtained for modalities, datasets, learning techniques, and languages showed the highest utilization of: combined modality (38.29%), Twitter dataset (38.29%), DL techniques (40.42%), and English language (74.46) in comparison to other modalities (text, speech/audio, and im-age/video), datasets (Facebook, Weibo, AVEC, Reddit, DAIC-WoZ, combined, and others), learning techniques (ML, both ML & DL), and languages (Code-mixed data, Urdu, Bangla, Chinese and Malay).

In conclusion, this chapter lays a compact foundation for developing deep learning frameworks using multimodal social media content for depression detection tailored to the challenges found in this chapter. The outlined research objectives provide a clear direction and set the stage for the novel models introduced in the subsequent chapters.

CHAPTER 3

A MULTIMODAL DEEP LEARNING-BASED FRAMEWORK FOR DETECTING DEPRESSION USING ENGLISH SOCIAL MEDIA CONTENT

3.1 Introduction

Social media has firmly established itself as an indispensable part of life for the majority of the population nowadays. With easy accessibility to internet services and mobile devices, people of all age groups indulge in social media websites such as Instagram, Twitter, Reddit, and so on. The content consumed by users on these platforms often takes a toll on their mental health, which can lead to disorders such as anxiety and depression. Surveys show that out of all age groups, teenagers spend the maximum time on social media with an average of 4.8 hours [128]. It has been observed that people who spend excessive time on social media are also the most vulnerable to depression and other mental health disorders, indicating a strong correlation between the two [104].

Mental health is not considered as important as physical health and is still not discussed openly because of social stigma or lack of acceptance due to understanding and awareness about it. Because of the ignorance of symptoms of mental disorders, it goes undetected for a prolonged duration, further intensifies and leads to mental distress. Therefore, the raw data posted on social media sites by a pool of users can be helpful for researchers and doctors to critically analyze and understand a person's mental state to diagnose depression at its early stage.

The advanced learning algorithms or models of Artificial Intelligence (AI), such as machine learning (ML), deep learning (DL), ensemble learning, transfer learning, etc., have the remarkable ability to understand complex patterns in data and to provide accurate results with high performance. ML techniques have proven

beneficial in healthcare due to their extraordinary capability of processing a large amount of data [41]. Also, Natural Language Processing (NLP) techniques can be efficiently utilized in sentiment analysis, allowing machines to understand words and text like humans accurately. Therefore, to analyze the severity of depression of a person based on their social media posts, NLP techniques are preferred, as shown by various studies [7].

Further, DL algorithms have automated the entire diagnostic process as they can efficiently deal with complex data without the intervention of human beings. These techniques allow the users to adjust the network parameters to enhance the prediction's accuracy [106]. Therefore, the incredible benefits of these computational methods have attracted the interest of researchers to utilize them for the practical analysis of persons with mental illness and depression issues.

Further objective of this chapter is to provide a unified approach for depression detection using English social media posts. So, the chapter is divided into two subsections. Section 3.2 that will explain the deep learning-based model to detect depression using multimodal English social media content for a binary class classification i.e., either depressive or non-depressive class and the proposed model is then compared with the existing state-of-the-art as well as hybridization of TL and ML based models but it lacks the severity aspect of depression which is then covered in the further subsection. Furthermore, Section 3.3 paves a path for multiclass classification using the deep learning-based framework on an English social media dataset that specifically lays the focus on the severity of the post i.e., model will classify the posts into 3 classes Severe depression, Moderate depression and No depression on a publicly available dataset.

3.2 Introduction to Deep Learning Based Depression Detection Model for Binary-class classification

The popularity of social media platforms is increasing exponentially due

to increased mobile devices and internet penetration, easy content creation and sharing facilities, growing social validation and feedback culture, etc. Various online applications, such as Instagram, LinkedIn, Facebook, Twitter, etc., allow users to interact and communicate with each other, thereby expressing their sentiments, feelings, and emotions on a particular topic [97]. Hence, these social media platforms provide an easy way for users to publicly share their opinions by responding to each other queries. However, the frequent use of social media for daily life offers significant advantages, but it involves serious drawbacks too. In this regard, state-of-the-art studies revealed that the high usage of social media platforms is directly proportional to increased depression and other mental disorders problems [104].

According to a World Health Organization (WHO) report published in the year 2019, almost a billion people were facing the problem of mental disorders, out of which 14% belonged to the youth age. It is observed that teenagers are the primary source of content generators as they are primarily active on social media platforms. Further, there is a significant rise in the suicide rate of youth between the ages of 15 to 29 years. As per one of the Hindustan Times (one of the leading news portals) reports [29], one student commits suicide every hour after posting messages like ending life or can't survive on social media platforms. In the recent past, there have been several incidents where people have even live-streamed their suicides on social media. One of the main reasons behind this drastic decision to end life was depression [31, 105].

An estimate by WHO also revealed that mental health problems are 2443 disability-adjusted life years (DALYs) per 10000 population and the suicide rate is 21.1, offering major economic loss in India [9]. The issues related to depression and anxiety have existed for a substantial amount of time, but their prevalence seems to have increased by 25% by the end of the first year of the COVID-19 pandemic [8]. The occurrence of COVID-19 has affected the world population to a significant extent, as a result of which their interest has shifted more towards online platforms to share feelings and thoughts.

The traditional process of diagnosis tends to be time-consuming, with

possibilities of false positives and false negatives. Therefore, an efficient and automated diagnostic process is required. To make the entire process more reliable and speedier, researchers have made remarkable progress in utilizing Natural Language Processing (NLP) for deep contextual understanding of human language, as well as automating the process using Artificial Intelligence (AI) models such namely Deep Learning (DL), Machine Learning (ML) transfer learning i.e. making use of existing models to tackle problems in hand, and ensemble learning, i.e., combining ML models for better performance.

3.2.1 Major contributions

This chapter aims to address the issues and cover the undiscovered gaps which are gathered through literature, a novel hybrid framework is proposed utilizing the concept of Transfer Learning (TL) and Deep Learning (DL) based on multimodal social media data for depression detection. The significant contributions of this chapter are as mentioned below:

1. As per the survey, no multimodal dataset exists for depression analysis. Therefore, firstly, a multimodal dataset is created and presented that contains the image and the corresponding text and emoticon data belonging to depressive posts.

2. Secondly, a unified, robust and hybrid framework is proposed that can work on any type of social media platform and data for depression detection using images as well as text by extracting more relevant information for a reliable diagnosis.

3. Thirdly, a hybrid TL and DL based approach combining Bidirectional Encoder Representations from Transformers (BERT) and Convolutional Neural Network (CNN) (BERT-CNN) is created to accurately classify posts as depressive and non-depressive by analyzing textual and image data separately.

4. Fourthly and lastly, the proposed approach is compared with existing state-of-the-art techniques along with other hybrid DL+ML approaches using various

performance metrics. Moreover, it was found that the proposed hybrid approach outperformed the other above-mentioned approaches by a significant margin.

3.2.2 Proposed Framework

This chapter presents a novel framework that can work on multiple platforms and can classify posts that are depressive and non-depressive based on multimodal data. The methodology implemented in this chapter is provided in Figure 3.1 and consists of different phases such as dataset collection, including text, emoticons, and images; preprocessing; classification; and performance evaluation. A detailed description of each phase is given below:

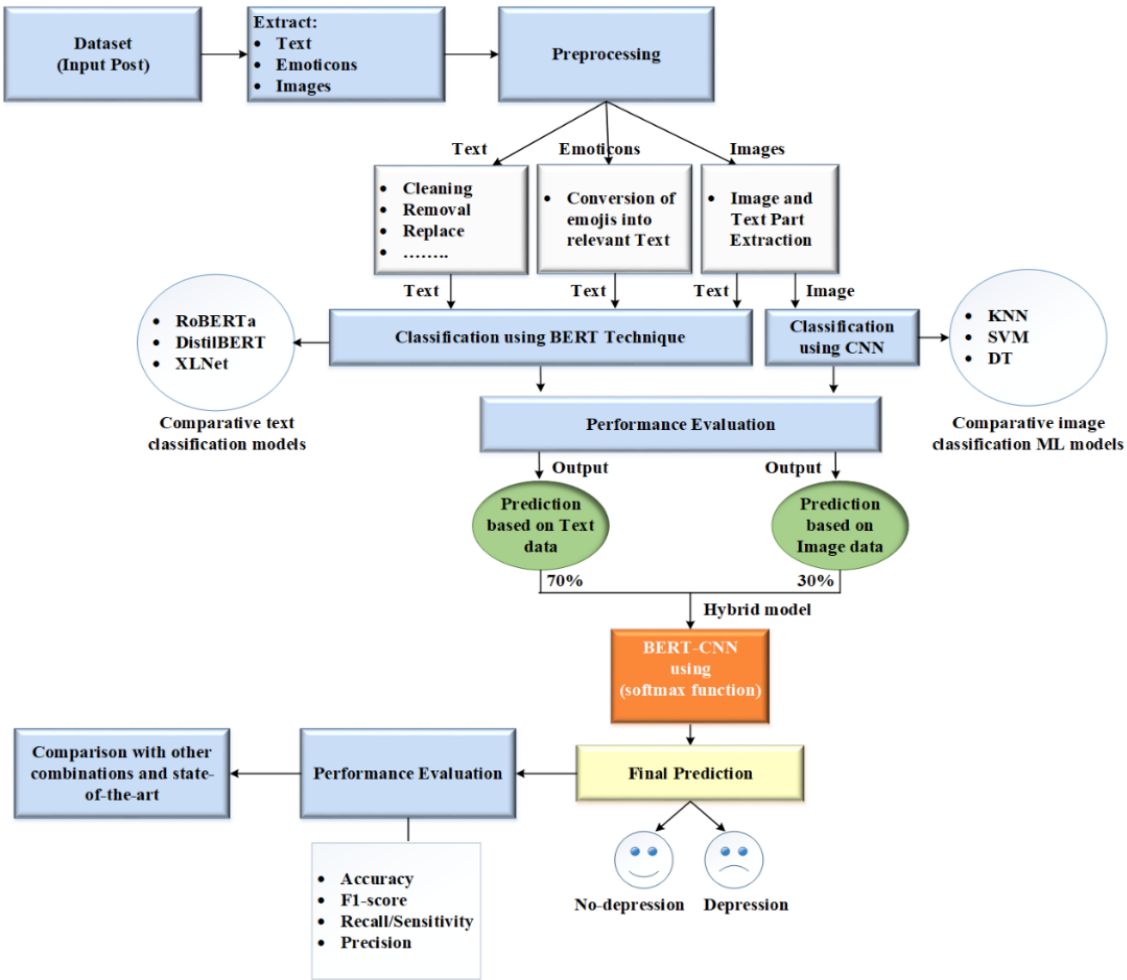


Figure 3.1: Proposed methodology for depression detection

3.2.2.1 Data Acquisition and Dataset Overview

A relevant dataset plays a crucial role in the depression detection process. The exploration of the literature revealed no public availability of a reliable multimodal dataset based on depression. Therefore, the significant contribution of this chapter is to present a dataset and the proposed novel DL hybrid approach to analyze depression from multimodal content, including text, emoticons, and images. One of the most popular social media platforms i.e., Instagram, has been used to gather the entire data. Those posts from Instagram users were downloaded and accessed, which were publicly available. The data is stored in CSV format consisting of the post, path of the image, and happiness index (HI) where HI=0 represents depressive post and HI=1 denotes non-depressive post. The entire data is annotated as depressive or non-depressive with the help of an expert psychiatrist from Futela Hospital. The collected data was labeled into 0 or 1 after in-depth analysis and evaluation by the expert.

A total of n= 10,295 posts, including text, emoticons, and images, were collected, out of which n= 3431 posts were depressive and n= 6863 were non-depressive, as shown in the following Figure 3.2. The validation of data is performed by the health experts categorizing posts into depressive and non-depressive. The data is also available on the GitHub [107]. Table 3.1 shows a sample list of data (text and images) used in this research work to classify depressed and non-depression individuals. In the table, the light grey shaded rows represent the text and emoticons data and their respective image data is presented in the rows shaded in dark grey color.

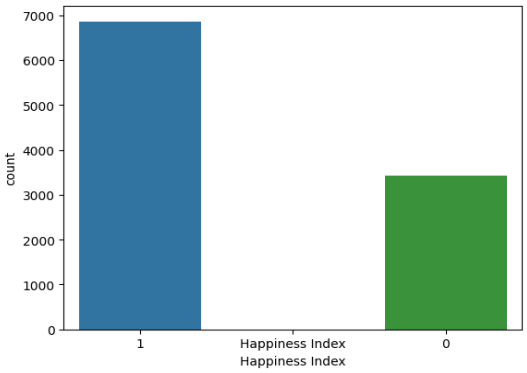
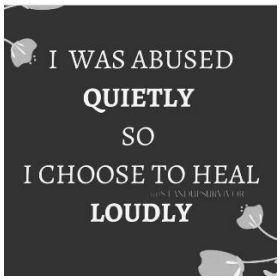
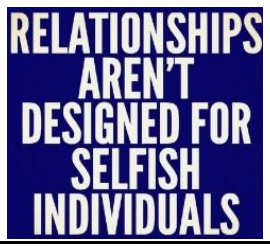
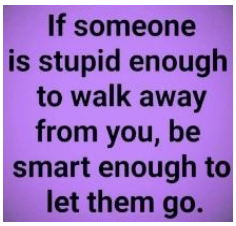
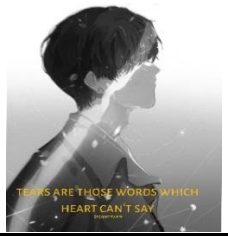
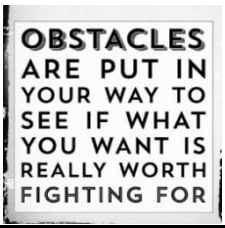
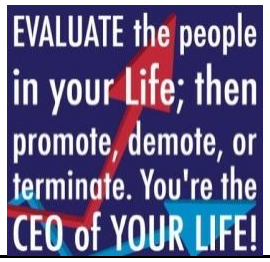
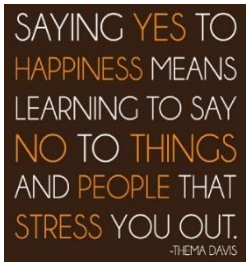


Figure 3.2: Dataset representations of non-depressive and depressive posts

Table 3.1: Examples showing depressive and non-depressive posts using multimodal data (text with corresponding images)

Sample Examples		
Text and Emoticons	**Image**	**Category**
A lot of people don't understand what it is like to be with a narcissist. SPEAK UP AND BE LOUD! You can help others. #narcissist #narcissim #narc #narcissticabuserecovery #narcissisticabuse #narcissistawareness #narcissistic #narcissistfree #narcissistquotes #depressed		Depressive
ABSOLUTELY...✌#bye #narcissist #narcissisticabuse #narcissism #whytolive		Depressive
It was always all about you. selfish piece of 💩. ✌#bye #narcissist #narcissim #sadforlife		Depressive
But those nights feels so heavy to sleep #sleepless #sad #broken		Depressive

Let them go 🤘... #byefelicia #nobigloss #dontgobacktowhatbrokeyou before #seeyoulateralligator #dontletthedoorhitya wherethegoodlord splitya		Depressive
Relatable 🖤? #lyftruths #sadquotespage #sadquotespage #sadedits ⚫ #sadedits #brokenquotes #sadlines #sadfeeling #sadline #lovefeelings #sadlove #brokenlove #hatelove #poetryworld #poemsbyme #depressionhelp #writer #depressededits 😳 🔫 #depression #relatablequotes #relatable		Depressive
Harsh Truth .. #broken #hate #depression #Nomorelife		Depressive
Fight for your dreams and never give up! #fighter #love #life #quote #obstacles #happiness #dreams #believe #selflove		Non-depressive
Enjoy your weekend and have fun! #happy #smile #fun #instahappy #funtimes #feelgood #enjoy #lovelife #laugh #weekend #selflove		Non-depressive

Absolutely love this quote!! #ceo #yourlife #happiness #freedom #picoftheday #instamood #gratitude #quote #selflove		Non-depressive
Your Past is Just a Story #moveon #goforward #letgo #past #history #live #fully #present #moment #future #destiny #create #story #life #unfiltered #truth #new #quote #selflove		Non-Depressive
Say Yes to Happiness #good #people #open #heart #give #take #learn #boundaries #yes #no #stress #relax #staycalm #letgo #accept #happiness #inner #peace #truth #mantra #new #quote #life #unfiltered #selflove		Non-Depressive

3.2.2.2 Preprocessing

Text plays a vital role and contributes majorly to comprising depression-related data as most of the users express their emotions via tweets, quotes, statuses, etc., frequently on social media. People often use text with emoticons to convey their feelings and sentiments that need to be correctly analyzed to perform accurate predictions regarding depression. Therefore, to preprocess the text, tokenization is done in which the given text is split into small units referred to as tokens [108]. Single words, entire sentences, phrases, etc. can be part of tokens. The process of tokenization involves the removal of characters such as punctuation, special characters, spacing, etc.

3.2.2.2.1 Text Processing Techniques

Several such techniques are employed in this chapter to clean and preprocess the text comprising of the following [109,110]:

• **Removing Stopwords:** Stopwords refer to the most commonly used words in the Stoplist that need to be refined or filtered as they don't have much significance. Since Stopwords are not of much importance, they are removed utilizing the standard list of Stopwords.

• **Removing Special Characters:** Some special characters have drawbacks in the way they are utilized. Further, the use of special characters can introduce unwanted noise in the data that can drastically affect the classification capability of a model. Therefore, apart from characters from A to Z and a to z, all other ASCII characters are eliminated from the data.

• **Discarding Numbers:** To represent a mood, emotions, thoughts, etc., of a person, numbers can't be utilized and should be discarded in order to have pure text characters. The presence of irrelevant data in the form of numbers can result in a high misclassification rate and decreased probability of accurate detection. Thus, the numbers are dropped from the text due to their negligible importance.

• **Removing Punctuations, Spaces, and Links:** While performing training of data, the use of punctuation can add irrelevant noise and lead to ambiguity. Each text can be treated as equal once the punctuation removal process is done. It enables the process to consume less memory and visualize the data. Also, the low-quality links are eliminated which were not contributing anything to the text information.

• **Removing Repeated Characters and Spell Righting:** Duplicated characters offer ambiguity in data and the short text abbreviations can lead to incorrect

classifications. Therefore, the characters that were repetitively present in the text and are not part of the English dictionary are removed. The short forms misguiding the correct meaning of the words are also discarded. Further, the words which have the possibility of being miswritten are replaced by their corrections. Spell correction is applied to the words using correct spelling suggestions based on Wikipedia, Oxford, dictionaries etc.

• **Conversion to Lower Case:** In many scenarios, it is suggested to convert the entire characters into lower case using an appropriate function on each word. Therefore, the idea is applied to convert the input text into a similar casing format using the lower () method. Once this entire text preprocessing was done, we moved ahead with the emoticons preprocessing.

3.2.2.2.2 Emoticons Preprocessing

Sometimes, people use many emoticons to express their moods, mental states, and feelings. This modality also forms an integral part from which a user can be classified into depressive and non-depressive based on their posted emoticons. Therefore, the emoticons are converted into text to extract relevant details and provide useful insights about the users. For example, emoticon such as '♥' is converted into 'black heart' text. The scores from emoticons can be utilized to analyze the mental state and emotions of a user. The value 0 is used to denote depression and 1 represents non-depressive post. Once the emoticons are transformed into the text, these are appended with the tokenized text due to their similar format.

3.2.2.2.3 Images Preprocessing

Like text and emoticons, image is also the most preferred modality, which, on proper visualization, can provide direct clues regarding the mental status of a person. The collected data consists of numerous images that also have the text data. Therefore, for efficient evaluation, extracting or separating the text part from the image is necessary. This extraction is performed using an "Optical Character Recognition"

(OCR) [111] technique, which has the feature to transform an image having printed or handwritten text. After extracting the text from all the images, it is processed following the same operations applied to tokenized textual data, and the image part is further processed using a DL technique.

3.2.2.3 Feature Extraction

Once preprocessing is done, the data is converted into the required format to perform further operations on it. The BERT model is applied to obtain the features from text [112] to efficiently use them for the classification of a user's post. The embedding tokens of sentences are fed as input to BERT, which is followed by sentence separation using special tokens. At last, the outcome of embedding for each token, known as the hidden state, is given as output by the BERT model. Similarly, to extract the features from images, CNN is employed that doesn't require heavy preprocessing and extracts the features automatically without any human intervention.

3.2.2.4 Classification

Classification refers to an ML procedure mainly supervised and used to categorize a given dataset into different classes depending on the training dataset. A sample known as a training example is provided as input to the model from which it learns and can classify the test example, which is an unknown data point. This section describes the entire process of classification using the proposed hybrid approach.

3.2.2.4.1 Proposed Approach

The chapter intends to perform classification by proposing a robust hybrid approach that can work on multimodal data. Therefore, to achieve the purpose of classification, the text and image parts are processed separately. The text data is evaluated using different TL techniques, including BERT, RoBERTa, DistilBERT, and XLNet. Further, the images are analyzed using DL as well as ML techniques to check their efficacy. The best models for both modalities are then chosen for further

operations and hybridized to make the final predictions. Therefore, the proposed model is created based on two highly-performing techniques to yield the best results. A detailed description of the techniques employed and the hybrid approach proposed is given in the following subsections.

3.2.2.4.1.1 Text Classification

i) BERT

Transfer Learning (TL) has recently improved text classification and other computer vision tasks. TL technique makes reuse of the pre-trained model to solve another task. A pre-trained model as the beginning point offers superb benefits despite creating a model from scratch. In this concern, the models based on transformers are being adopted in a broader range by researchers as they don't require labeled data and can be quickly accelerated with the help of GPUs [113]. Therefore, in this research, one such popular model known as BERT is utilized to perform text classification to detect depression. This model is a pre-trained model that learns the text sequence-wise. Thus, BERT, being Bidirectional, processes the text taken as input in both directions compared to other traditional methods that analyze text in only one direction. Due to this property of BERT, high performance can be achieved by capturing many contextual details.

The working of BERT involves embedding input tokens with the help of an embedding layer as the first step, then using a transformer encoder consisting of various self-attention layers to capture distinct parts of the input sequence. Finally, the output of the transformer encoder, contextual token embedding, is given as input to the output layer to make final predictions [114]. The final output generated by the model is the class label, i.e., 0 or 1, representing depressive and non-depressive posts. There are different versions of BERT, but BERT-Base is preferred in this chapter as it is affordable, smaller, and computationally efficient. The basic structure of BERT-Base also explains the architecture of a single encoder, as shown below in Figure 3.3, where several encoders are stacked over each other. The used BERT version consists of 12 layers of encoders arranged in a stacking pattern, i.e., 12 attention heads and 110 million parameters. The number of parameters and the attention heads in these models

increases with the number of layers. During the training process, masking is performed on the portions of tokens taken as input. Then, the model predicts the masked tokens by learning the association among words and their respective meaning in the context of the input sequence.

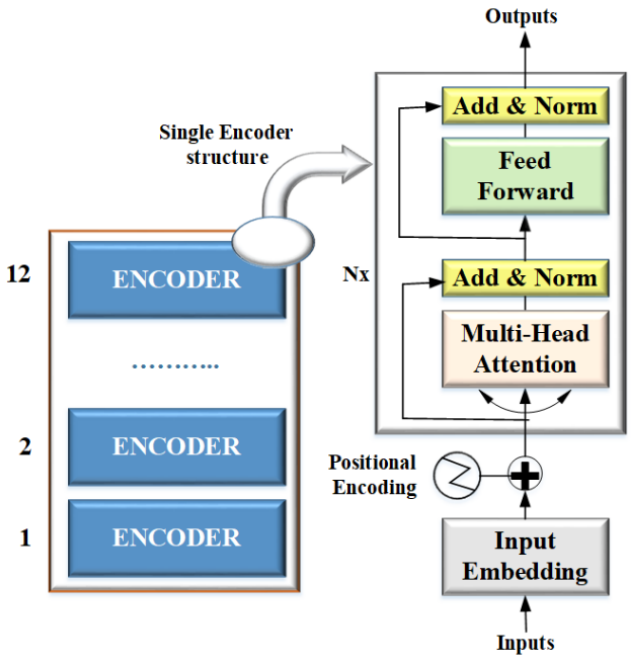


Figure 3.3: Architecture of BERT-Base Model

ii) Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa is another version of the BERT model invented by the AI researchers at Facebook and utilizes a self-attention procedure to evaluate the input sequences. A significant difference between BERT and RoBERTa is that the latter can be efficiently applied to larger datasets and uses an advanced training process. The model can learn generalizations and complex work patterns due to the dynamic-making method used by RoBERTa [115]. Another difference is that RoBERTa uses a long training learning rate and needs to train for longer sequences. RoBERTa neither requires defining tokens belonging to a particular sentence nor using token ids. The architecture of RoBERTa is very much like the BERT model but with some modifications of important hyperparameters.

iii) DistilBERT

The distilled version of BERT is referred to as DistilBERT. These models are smaller, faster, and cheaper, in which the BERT model size is reduced to 40% by knowledge distillation. DistilBERT, with the help of the used distillation procedure, approximates the more extensive neural network with the smaller one [116]. This concept is similar to posterior approximation, which uses the Bayesian statistics theory. This version of BERT doesn't allow selecting input positions and using token ids like RoBERTa.

iv) XLNet

XLNet is the newly developed unsupervised language that uses Transformer-XL as the base model and has improved significant performance in textual classification. It utilizes an improvised mechanism for training and possesses enormous computational power compared to BERT. The improvement in the training is made by using modeling referred to as permutation language modeling. The main difference between XLNet and BERT is that in the former, the predictions for all the tokens are performed randomly [117]. In the case of BERT, predictions are only made for those tokens which are masked. Also, the concept behind XLNet contrasts traditional language models where sequential order is followed to make predictions on tokens.

3.2.2.4.1.2 Images Classification

i) DL-based Designed CNN

To differentiate user's post with depression and no depression based on image modality, CNN is used, which is a type of DL technique and is most widely preferred in text analysis. One of the unique properties of CNN is that the parameters, such as batch size, drop out, layers, etc., can be easily varied to achieve the objective function with high accuracy. These models have a layered architecture and are computationally very efficient. Different layers contribute to the building of CNN, such as convolutional, pooling, flattening, fully-connected, etc. Also, the output of the

model summary is presented in the following Figure 3.4.

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_9 (Conv2D)           (None, 254, 254, 32)      896

 max_pooling2d_9 (MaxPooling  (None, 127, 127, 32)     0
 2D)

 conv2d_10 (Conv2D)          (None, 125, 125, 64)      18496

 max_pooling2d_10 (MaxPoolin  (None, 62, 62, 64)       0
 g2D)

 conv2d_11 (Conv2D)          (None, 60, 60, 64)        36928

 max_pooling2d_11 (MaxPoolin  (None, 30, 30, 64)       0
 g2D)

 flatten_3 (Flatten)         (None, 57600)             0

 dense_6 (Dense)             (None, 128)               7372928

 dropout_3 (Dropout)         (None, 128)               0

 dense_7 (Dense)             (None, 1)                 129

=================================================================
```

Figure 3.4: Output of Model Summary

In this chapter, following the model's basic structure, the adjusted CNN is created to perform such classification with high performance, as shown in Figure 3.5. Various layers added at different positions in the modeled CNN comprised of the following layers [118, 119]:

- **Convolutional Layer:** The initial extraction of features from the input data is the responsibility of the first layer of CNN, namely the Convolutional Layer. This layer serves as the base for other layers and, therefore, is considered the building block of CNN. This layer makes use of small matrices with size 2×2, 3×3, 5×5 or 7×7 known as filters or kernels, which is convolved with the input image by applying dot product. The output generated from this operation is the feature map given as input to the next layer. In the proposed architecture of CNN, 3 convolution layers

are used at the 1st, 3rd, and 5th positions with different numbers of filters.

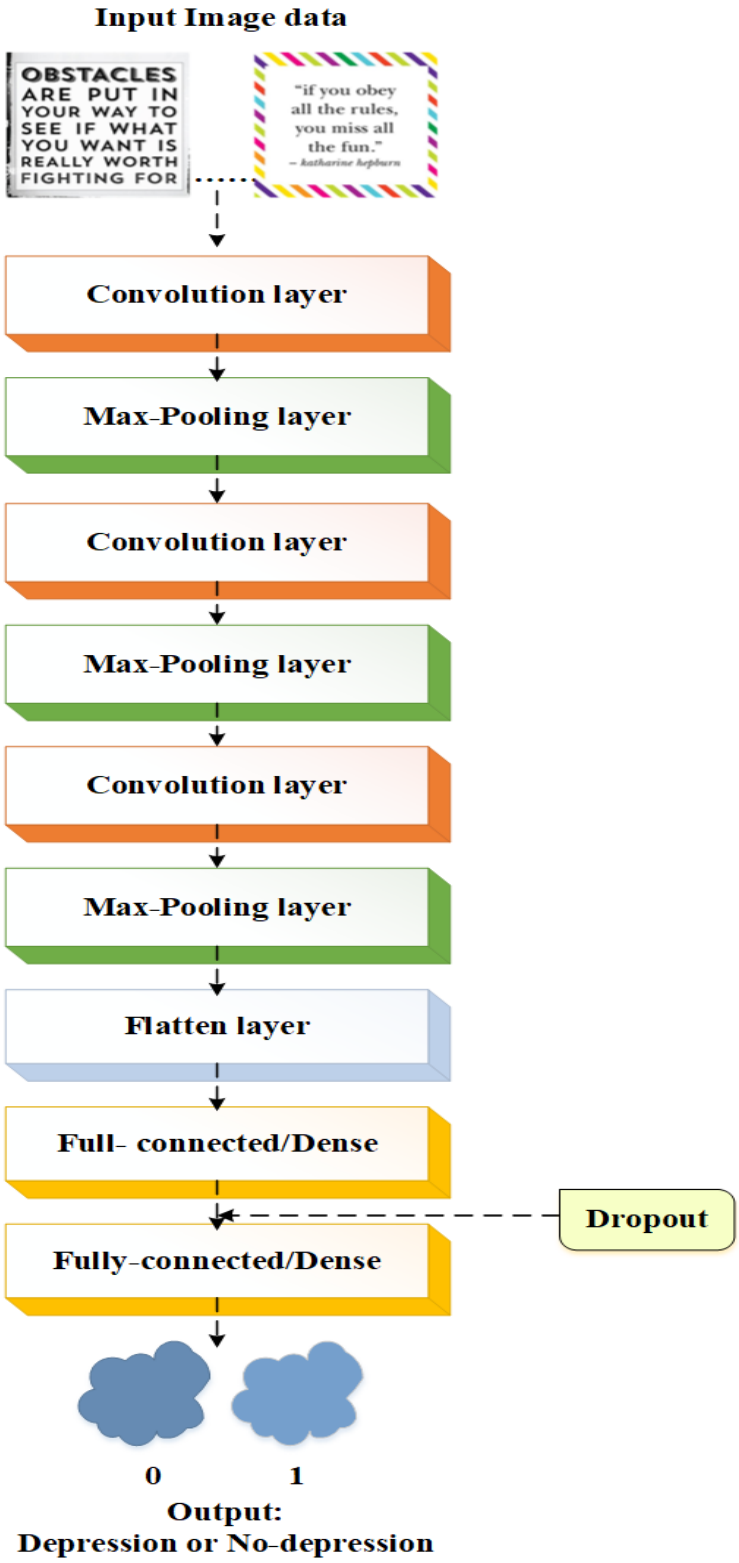


Figure 3.5: Designed Architecture of CNN

- **Max-Pooling Layer:** A pooling layer is applied to reduce the input size by decreasing its dimensionality. It eliminates the less relevant features from the actual data and makes the entire process computationally fast. This layer is usually inserted between the convolutional layers and only makes improvisations/updations in the required data information. The pooling layer also can solve the overfitting problem and build a new set of pooled feature maps from the old one. In the current model, 3 such layers are inserted at the 2^{nd}, 4^{th}, and 6^{th} positions.

- **Flatten Layer:** After the convolutional and max-pooling layer, the flatten layer is added to convert the outputted feature maps into one-dimensional vector form to apply further operations. The result of this layer is then sent to the dense or fully-connected layer as input. In the proposed model, 1 fatten layer is inserted at the 7^{th} position.

- **Dense or Fully-connected Layer:** This comprises the last CNN layer responsible for generating the classified output. The fully-connected layer takes the input from the flattened layer and performs mathematical computations to perform accurate classification. The output obtained from this layer is then sent to a logistic function that transforms the outcome into a probability score or label belonging to a particular class. Here, the softmax activation function is adopted to obtain probability distribution from the vector of values and can effectively perform multi-class classification represented in equation (3.1). Two fully-connected layers are used at the 8^{th} and 10^{th} positions in the designed network.

$$output = act(dot(it + KW) + bi) \tag{3.1}$$

where *dot* represents dot product of weights and input *it* denotes input data, *KW is* the weight data, and *bi* denotes biased value.

- **Dropout Layer:** The dropout layer is added at the 9th position in the model architecture and can overcome the overfitting problem. This layer provides stability to the model and makes its processing faster. Due to the addition of this layer, a model can efficiently learn the relevant features of the image. The different parameters are set to perform the classification using designed CNN, such as learning rate= 0.01, dropout = 0.5, momentum = 0.09, batch size = 15, and epochs = 300.

ii) ML-based Models

a) KNN: KNN is the widely preferred ML classifier based on a supervised platform where a grouping of data points is performed using proximity. This algorithm assumes the minimum distance between the samples belonging to similar classes. In KNN, firstly, the value of 'K' is set and then a metric is used, such as Euclidian distance, to compute the distance among the 'K' number of neighbours. Based on the value of 'K', the data points with minimum Euclidian distance from classes are selected and the new data point is assigned to the class having the majority of selected data points. KNN is simple to implement and is also known as a lazy learner as it doesn't use a training dataset for learning immediately but stores the dataset to use during the classification process [120]. The distance calculation between two points using the Euclidean distance (Euc) measure can be done using equation (3.2).

$$Euc(u,v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \ldots + (u_s - v_s)^2} \qquad (3.2)$$

Here, u and v are the samples that need to be measured with the help of s characteristics. One of the significant hurdles in KNN is to find out the optimal value of 'K' on which the selection of neighbour depends. Therefore, to avoid ties, it should be chosen in odd numbers.

b) SVM: SVM is the supervised ML algorithm that aims to determine an optimal hyperplane or line that can perfectly classify the data points in N-dimensional space. The hyperplane that provides the maximum distance or margin between the data points of both samples is chosen. Mostly, the hinge loss function is used that performs well in maximizing the margin and can be represented as in equation (3.3)

$$h(u, v, f(u)) = \begin{cases} 0 \; if \; v * f(u) \geq 1 \\ 1 - v * f(u), else \end{cases} \tag{3.3}$$

In SVM, to achieve a low classification error rate, keeping the maximized margin between two classes is mandatory. The role of the kernel in SVM is very significant, which converts the low input space into higher dimensional space efficiently. This kernel trick makes the SVM more accurate, robust, and flexible [121].

c) Decision Tree (DT): DT is a tree-structured supervised learning classifier where the dataset features are represented by internal nodes. These trees' branches and leaf nodes indicate a decision rule and the respective outcome. DT starts from the root node and keeps on expanding to branches to perform the operations on the dataset, thus forming a hierarchal tree structure. The matching of the record and root attribute is done until the leaf node of the tree is encountered. Two techniques, namely Gini Index and Information Gain, are mostly preferred to select the root attributes. DT can perform at high speed on large datasets but suffers from the problem of overfitting [122]. Therefore, an ensemble-based technique known as RF offers a solution to the issue in DT by combining multiple DTs [123]. Consider D_n as the data present at node n with s_n samples and t_n thresholds. The algorithm for the classification tree can be presented in equation (3.4) as

$$G(D_n, t_n) = \frac{s_n^{Left}}{s_n} H(D_n^{Left}(t_n)) + \frac{s_n^{Right}}{s_n} H(D_n^{Right}(t_n)) \tag{3.4}$$

where H is the left and right impurities measurement at node n and the number of instances is denoted by s_n.

3.2.2.4.2 Proposed Hybrid BERT-CNN Approach

The gaps in the existing literature, including the lack of a multimodal dataset, a robust multimodal model, simultaneous evaluation of text and images, and much less use of hybrid advanced learning platforms, encouraged this research to propose and present a reliable strategy to deal with multimodal data on social media platforms. There are different models that can either work on text or image data for depression analysis but lack the simultaneous evaluation of text and image data. Due to the high variability in the content posted by users on social media, it is necessary to develop a unified model that can recognize whether a person is suffering from depression or not based on multimodal data. Therefore, this chapter presents a hybrid model combining TL and DL platforms to deal with text and the image simultaneously.

The performance evaluation of all the applied techniques for text and image presented in the previous sections showed the highest performance with BERT and CNN. Therefore, these techniques are selected to create a hybrid approach by combining them to make the final prediction based on prior individual predictions. The text data is classified using Bidirectional Encoder Representations from Transformers (BERT) and for images; CNN is adopted due to its significant features and higher accuracy than ML techniques. Afterward, a hybrid model is designed by the combination of these two applied techniques, i.e., BERT+CNN=hybrid approach, for efficiently detecting depressive and non-depressive users. A schematic diagram showing the proposed hybrid BERT-CNN concept, designed to overcome research gaps, is presented in the following Figure 3.6.

In depression detection, textual data has been given more preference among all the modalities due to its more accurate results. Therefore, in this chapter, the separate predictions made by BERT and CNN on text and image data are taken as input in the ratio of 70:30, i.e., 70% text and 30% images. This research considered several combinations of this text:image ratio, such as 50:50, 60:40, 80:20 etc.; however, the best results are achieved with the above-mentioned 70:30 ratio.

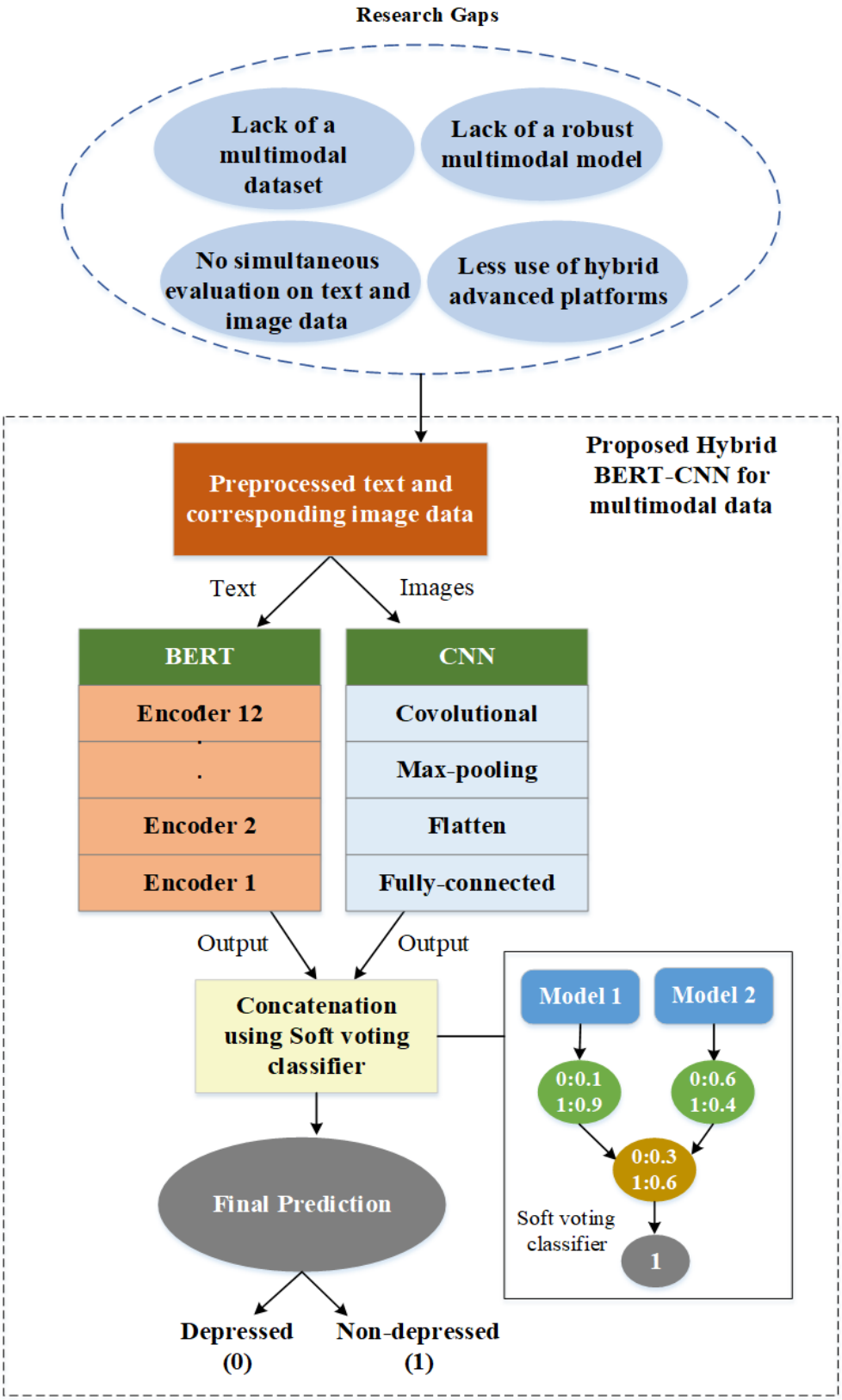


Figure 3.6: Concept of Hybrid BERT-CNN

As shown in Figure 3.6, the collected data is first preprocessed using the procedures and techniques elaborated in previous sections. The preprocessed text data with corresponding images is then fed separately as input to BERT and CNN models. The BERT model comprises certain transformer encoder layers built to process the text data and generate the output by applying certain mathematical operations. Similarly, the image data is passed through several CNN layers by increasing the number of abstractions to capture useful information about the image features. The fully-connected layer of CNN generates the output, classifying the image into depressive or non-depressive. A thorough structure of the designed BERT and CNN is given in subsequent sections.

Finally, the outcomes of both models are fused using an aggregation technique known as soft voting classifier to generate the final class prediction or label. Based on the probability of predictions, this voting procedure works to hybridize the predictions from BERT and CNN, thus forming a multimodal text and image model. In Soft voting classifier algorithm, each model assigns a probability value to each class, indicating that a specific data point refers to a particular class. In our case, there were two classes i.e., 0 and 1, representing depressed and non-depressed posts. The prediction probabilities by the models for each class are summed up and the final prediction is simply the class yielding the highest probability value. Therefore, following this procedure, the hybrid model is designed and the final results are generated to detect the depression accurately.

3.2.3 Experimental Results and Analysis

The entire training process was performed on Anaconda Navigator in Python programming platform with GPU system using the TensorFlow deep learning tool and Keras library. The experimentation was performed for three models adopted in this work: (i) BERT for textual data (ii) CNN model for image data and (iii) a hybrid approach combining BERT-CNN for multimodal data. Each of the models is evaluated

one by one and their respective comparison is also performed with other TL and ML techniques to find the best one. The performance evaluation metrics used are accuracy, sensitivity/recall, precision and F1-score which is explained in Chapter 2 Section 2.6.

To deal with text data, BERT is applied, which yielded an accuracy of 97.31%, Sensitivity or recall of 97.21%, precision of 97.14%, and F1-score of 97.13%, respectively. Further, the BERT model is analyzed by comparing it with other versions of BERT, including RoBERTa, DistilBERT, and XLNet. The results obtained reveal: 81.26% of accuracy, 75.31% of Sensitivity, 74.37% of precision, and 74.19% of F1-score with RoBERTa; 81.13% of accuracy, 76.42% of Sensitivity, 76.25% of precision, and 76.34% of F1-score with DistilBERT; and 12.49% of accuracy, 36.37% of Sensitivity, 54.46% of precision and 25.45% of F1-score with XLNet. The overall comparison based on evaluation metrics demonstrates the highest performance with the BERT model compared to other TL models in the classification of textual data. The result values obtained for all the models employed for textual data are provided in the following Table 3.2 and are graphically shown in Figure 3.7.

Table 3.2: Results for classification of text data into depressive and non-depressive

Model	**Accuracy**	**Sensitivity/Recall**	**Precision**	**F1-score**
BERT	97.31%	97.21%	97.14%	97.13%
RoBERTa	81.26%	75.31%	74.37%	74.19%
DistilBert	81.13%	76.42%	76.25%	76.34%
XLNet	12.49%	36.37%	54.46%	25.45%

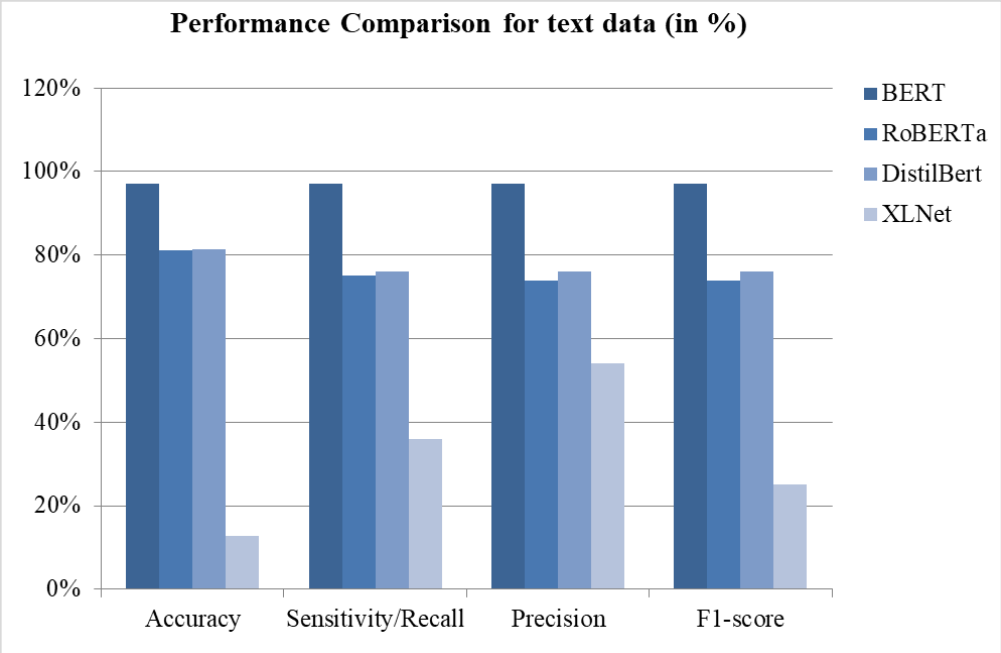


Figure 3.7: Graph showing results for Text data classification using different techniques

After the evaluation of textual posts, the classification of image data is performed with the help of the designed CNN model due to its tremendous properties in image-related tasks. The results achieved using the applied DL-based CNN model reflect the significant performance by attaining an accuracy of 89.42%, Sensitivity or recall of 83.26%, precision of 86.31%, and F1-score of 83.15%, respectively. The experimentation is extended by considering the ML models in this process, where KNN, SVM, and DT were employed to check their efficacy. The accuracy rates of 73.35%, 81.43%, and 81.41%, the Sensitivity of 73.26%, 81.29%, and 81.42%, the precision of 79.19%, 82.38%, and 82.38%, and the F1-score of 73.48%, 81.37%, and 81.45% were obtained using ML techniques i.e. KNN, DT and SVM. The comparative analysis of results obtained using DL and ML models shows that the efficiency of the CNN model outperforms the ML techniques by achieving the highest accuracy rate. The result values for different parameters attained for CNN and ML methods are given in Table 3.3 and are graphically presented in Figure 3.8.

Table 3.3: Results for classification of image data into depressive and non-depressive

Model	Accuracy	Sensitivity/Recall	Precision	F1-score
CNN	89.42%	83.26%	86.31%	83.15%
KNN	73.35%	73.26%	79.19%	73.48%
DT	81.43%	81.29%	82.38%	81.37%
SVM	81.41%	81.42%	82.38%	81.45%

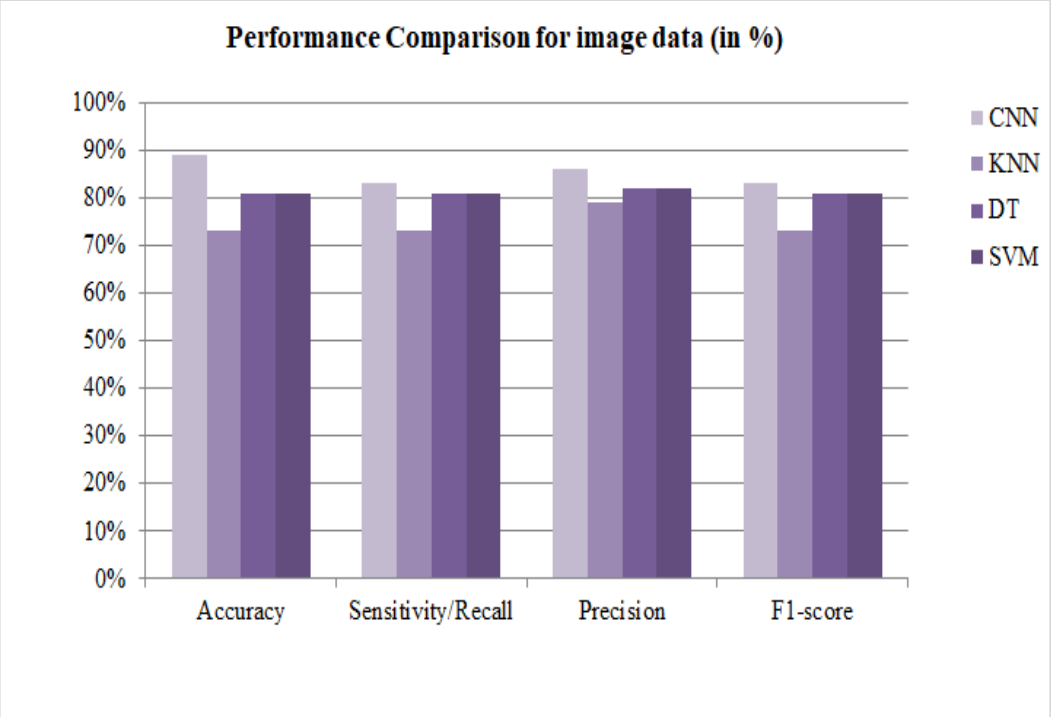


Figure 3.8: Graph showing results for Image data classification using different techniques

The graph in Figures 3.7 and 3.8 reveals the highest performance using BERT and CNN for textual and image data. Therefore, to take advantage of both techniques in this study, a hybrid approach is proposed by combining BERT and CNN, i.e., BERT-CNN. To evaluate the proposed BERT-CNN model, the final prediction is

made with 70% weightage given to textual data while the remaining 30% weightage is given to image data due to more authenticity of the textual information. The results achieved using the proposed hybrid approach are shown in Table 3.4 and reveal the best performance. Further, to validate the results, the proposed BERT-CNN is comparatively evaluated with other combinations such as BERT-KNN, BERT-SVM, and BERT-DT. The performance evaluation metrics are computed for each model and the results achieved proved the efficacy of the proposed approach by yielding an accuracy rate of 99.31% in comparison to other methods with accuracy of 58.30%, 58.20%, and 65.10%, respectively, for BERT-SVM, BERT-KNN, and BERT-DT.

Table 3.4: Results for classification of combined data into depressive and non-depressive using the proposed method

Model	Accuracy	Sensitivity/Recall	Precision	F1-score
BERT-CNN (Proposed)	99.31%	97.32%	99.59%	99.16%
BERT-SVM	58.30%	51.38%	56.31%	52.47%
BERT-KNN	58.20%	51.41%	56.23%	52.43%
BERT-DT	65.10%	58.38%	63.28%	59.38%

Figure 3.9 shows that the highest performance values are attained by applying the proposed technique, and BERT-KNN shows the lowest results. Therefore, this study evaluated each algorithm separately and then in combination to cross-validate all the results. Firstly, the text classification techniques are analyzed by evaluating them individually. Among all, BERT has shown the highest accuracy for text classification into depressive and non-depressive categories. Afterward, a DL-based CNN method is evaluated individually in comparison with four ML algorithms, out of which CNN has outperformed the other ML methods in detecting depressive and non-depressive users from images. Finally, the best-performing models, i.e., BERT and CNN, hybridized to make a unified and robust model (BERT-CNN) that

can work with every type of data efficiently.

Further, the comparison of the proposed hybrid approach attained the highest results compared to other hybrid models. The results obtained in the above tables demonstrate the efficiency of CNN in extracting the most relevant features from the data by increasing the level of abstraction and adjusting the model's parameters. Also, the unique feature of the BERT model to process data in bidirectional allows for extracting more contextual details and improving the model's overall performance. The increase in accuracy rate is also observed when BERT and CNN are used in combination to classify multimodal data to detect depression. Therefore, the best accuracy is achieved when deep feature extraction capabilities of the CNN model are combined with the BERT model for an efficient image and text classification.

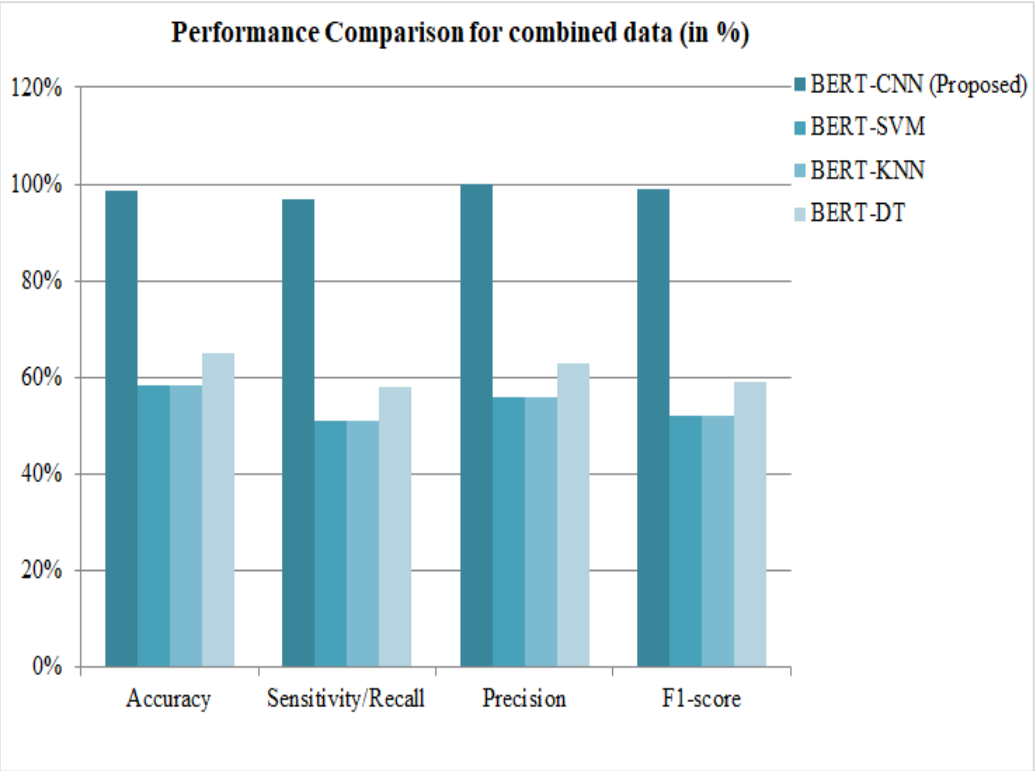


Figure 3.9: Graphical analysis of results using the proposed method on multimodal data

3.2.3.1 Statistical Analysis

The results achieved in this chapter showed the highest performance with the proposed approach, which is then validated by using the significantly known statistical test named as Friedman's Rank Test (FRT). To perform statistical analysis using FRT in order to evaluate performance of different models and the proposed approach, two other datasets i.e., [124 & 125] are also considered. Both datasets contain depression-related data, which is the area of focus of this research study. The different accuracy rates achieved by the employed and the proposed approaches on the considered datasets is given in Table 3.5 below, upon which the FRT is performed. The FRT helps to validate the substantial differences between the models. An evaluation is performed based on ranks after forming the null hypothesis. Consequently, a null hypothesis states that all the methods exhibit the same accuracy rate and no considerable visible difference. Moreover, the FRT allocates the highest rank to the model with the highest accuracy rate and the lowest to the model with the lowest accuracy rate. The test results attained a p-value of 0.0421, lower than the significance level, i.e., $p<0.05$. The test results showcased the fact that there is a significant difference in model performance on different datasets, subsequently rejecting the null hypothesis. The highest rank, i.e., rank 4, is attained by the proposed model, representing the highest performance and the lowest rank is attained by the BERT_SVM model. Table 3.6 below shows the FRT results, providing the average rank of methods used and the respective p-value.

Table 3.5: Accuracy achieved by models on different datasets

Datasets/Models	Proposed	BERT_SVM	BERT_KNN	BERT_DT
Proposed	0.9900	0.5830	0.5820	0.6510
[124]	0.9030	0.6250	0.6740	0.7330
[125]	0.9280	0.6690	0.6920	0.7540

Table 3.6: Results of FRT

Rank	1st	2nd	3rd	4th	p-value
Model	BERT_SVM	BERT_KNN	BERT_DT	Proposed	0.0421
Average rank	1.3333	1.6667	3	4	

3.2.3.2 Comparison with state-of-the-art studies

The proposed approach is compared with the current state-of-the-art studies to analyze the robustness and efficacy. Accuracy is utilized as the performance evaluation measure, which has been employed due to its high adoption in all the research works for comparison purposes. Table 3.7 provides an insight into this comparison that can be visualized graphically in the following Figure 3.10. Ramalingham et al. [126] used Twitter data in the form of images and videos to analyze depression with the help of SVM. They achieved accuracy rates of 82.2% and 70.5% for predicting males and females with depression.

Similarly, another study by Wang et al. [95] employed BERT, SVM, NB, LR, CNN, and LSTM on text, tags, and emoticons data acquired from the Weibo platform. The highest accuracy rate of 75.6% was attained by using the BERT model, thus showing its efficacy in depression prediction. The analysis of depression using Electroencephalography (EEG) signal data was performed by Li et al. [127]. The EEG data was fed as input to SVM, CNN, an ensemble of deep forest, RF, and KNN, and the highest results were obtained with LSTM, i.e., 83%. This chapter considered evaluating multimodal data (text, emoticons, and images), which previous studies have not utilized much. The research aims to propose a robust approach that can be used for all types of data to detect users with depression and no depression. The data in Table 3.7 indicates that the proposed hybrid BERT-CNN approach has outperformed the current state-of-the-art with an accuracy of 99.31%. Therefore, the proposed model

has a high potential to predict depression with optimum performance.

Table 3.7: Comparison of the proposed approach with previous studies

Author	**Year**	**Modality**	**Classifier**	**Dataset**	**Average Accuracy**
Vandana et al. [54]	2023	Text and audio data	Textual CNN, audio CNN, LSTM and Bi-LSTM	DAIC-WoZ	88% (with Bi-LSTM)
Gupta et al.	2022	Text	SVM, DT, KNN, LR and LSTM	Kaggle and Twitter	Highest with LSTM = 83%
Wang et al. [93]	2020	Text, tags and emoticons	BERT, SVM, NB, LR, CNN and LSTM	Weibo	Highest with BERT= 75.6%
Ramalingam et al. [126]	2019	Twitter like data (videos, pictures)	SVM	Weibo posts	82.2% (for males) 70.5% (for females)
Li et al. [127]	2019	HydroCel Geodesic Sensor Net	SVM, CNN, Ensemble of deep forest and SVM, RF, KNN	EEG features	Highest with ensemble = 89.02%
Proposed Study		**Text, images and emoticons**	**BERT (text), CNN (images), SVM, hybrid BERT-CNN (proposed)**	**Instagram and other social media posts**	**Highest with proposed BERT-CNN = 99.31%**

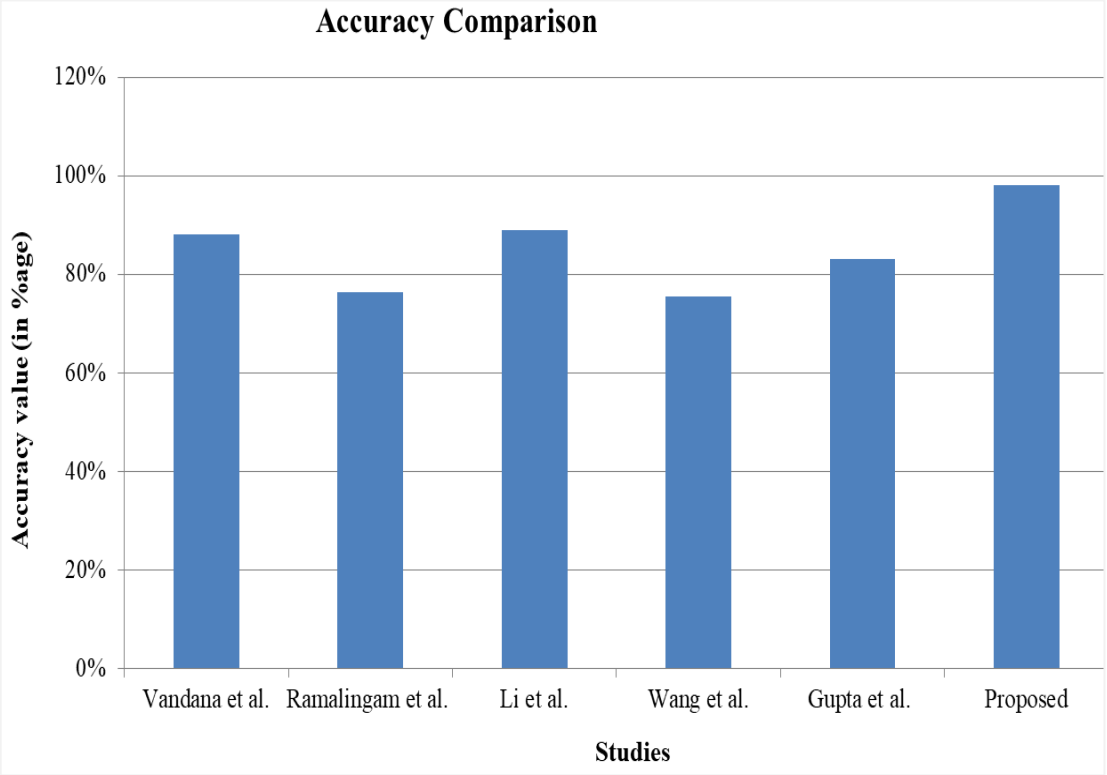


Figure 3.10: Comparison analysis of the proposed approach with state-of-the-art based on the accuracy

3.3 Introduction to Deep Learning Based Depression Detection Model for Multi-class classification

Nowadays, Depression, which impacts millions of people across the globe, ranks as one of the forms of mental illness [129]. Stress, low moods, nervousness, insufficient sleep, problems with eating, regret sensations, thoughts about suicide and attempts, along with other indications are among the most frequently experienced ones caused by depression [54]. More than 350 million individuals worldwide i.e. 4.4% of the global population are reported to suffer from depressive disorders, based on data from the World Health Organization (WHO) [130]. Further, by 2030, it is projected that depression will appear as the second leading reason for disability globally [131]. Depression, being a negative disorder, affects people at several severity levels,

including early, moderate, and severe. Identifying it early may prevent serious adaptation problems for the concerned person [132].

Several rating scales as well as questionnaires are employed in the traditional methods used in clinical settings for determining the severity of depression. These theoretical tools demand considerable amounts of time and primarily depend on input from patients and medical guidance from experts. A faulty and unreliable assessment may result from inadequate questionnaire filling and a lack of understanding of the subject matter. Thus, researchers are drawn to utilize automated cutting-edge techniques for facilitating accurate depression analysis [133]. Over the recent years, Machine learning (ML) and Deep Learning (DL) have made immense progress in the detection of affected mental health by providing standardized evaluation and automated analysis that could be helpful to lighten a certain amount of stress for clinicians [134].

Various social media online platforms including Twitter, Facebook, Instagram, etc. have become the prime source where many people share their emotions, thoughts, and feelings on a regular basis. Thus, a large amount of such data can be carefully analyzed to understand the behavior of a person to a large extent [135]. Focusing on text data available on social platforms, this chapter proposes a hybrid model incorporating transfer learning (TL) i.e. BERT and DL i.e. RNN models to detect and classify depression at three severity levels. Using a combination approach, rather than a single model, may improve the model's overall performance.

3.3.1 Proposed Framework

The proposed methodology comprises different stages that initiate from the dataset and end with a performance evaluation, as shown in Figure 3.11.

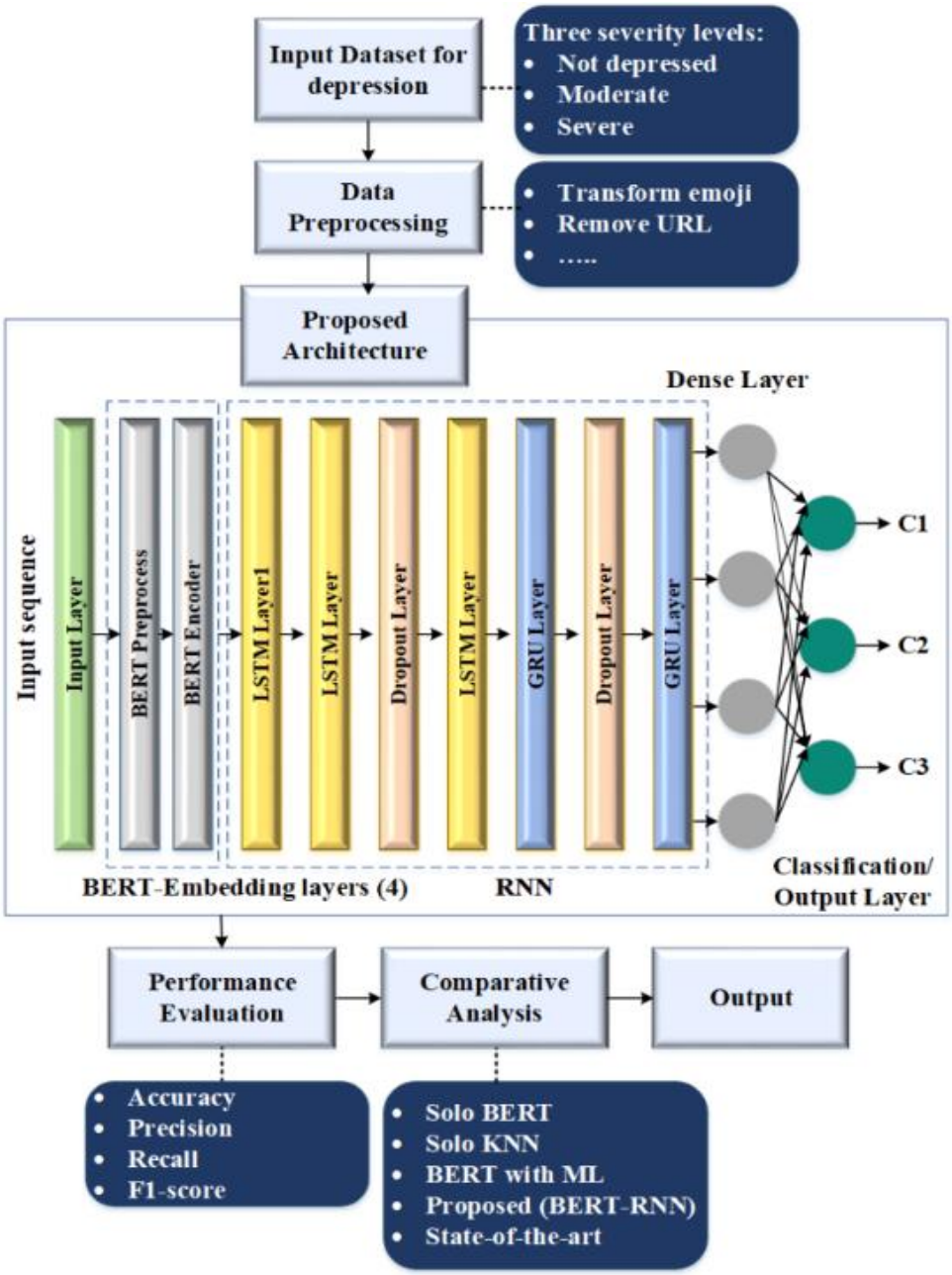


Figure 3.11: Proposed Framework for depression classification at different severity

3.3.1.1 Dataset

The dataset created by Durairaj et al. [136] is utilized in this study for experimentation. It consists of posts from Reddit in the English language, where all of them is associated with some categories or labels i.e. not depressed, moderate, and

severe. The label "not depressed" indicates no depression sign at all whereas "moderate" and "severe" labels represent slight depressive and high depressive symptoms and behavior. The entire dataset is split into training (n=8891), development (n=4496), and test (n=3245) sets. While using ML and DL approaches, it is good to have large training sets as compared to validation and test sets. This is because the efficacy of these techniques directly relies on the number and variation in samples during training. A brief description of the considered dataset is given in Table 3.7.

Table 3.8: Dataset Description

Label	Example	Train Set	Dev Set
Not Depressed	Happy New Year everyone……	1971	1830
Moderate	It is the worst feeling anyway…..	6019	2306
Severe	I just want to fall asleep forever…….	901	360
Total Samples	--	8891	4496

Preprocessing is a group of procedures that transform unprocessed data into useful forms to carry out the next steps with high accuracy [137]. Thus, to boost the worth of the raw data, firstly, a search for all the emoji present in the data is performed and then these are transformed into text form. Similarly, URLs or Links that did not offer anything of significance to the textual content were deleted.

3.3.1.2 Classification

DL models are known to be enhanced by employing the TL technique. The training of a neural network is first done as a linguistic model utilizing an extensive and complete set of data in the preliminary stage of the TL designing, which is usually named as semi-supervised training. This is then followed by supervised training, in which the model is trained to employ adequately labeled training information set [138]. In this section of this chapter, a novel approach combining both TL and DL techniques,

namely BERT and RNN, is proposed to perform text analysis and classification tasks. A brief overview of the techniques used is provided as follows:

3.3.1.2.1 Proposed BERT-RNN Hybrid Model

3.3.1.2.1.1 BERT Model

Many pre-trained models exist, but each of them suffers an identical issue i.e. their unidirectionality, which limits their abilities. As a consequence, a particular type of transformer network referred to as Bidirectional Encoder Representations from Transformers (BERT) came to light [139] that works by processing the text that is inputted in both directions. This ability of BERT can result in improved model performance. In this work, In this study, Bert-base architecture is employed due to its computational efficiency, compact size, and simplicity. It has 110 million parameters, 12 self-attention heads, and 12 layers of transformer encoder. WordPiece embeddings are employed to pre-train BERT. Each sentence in BERT's input encounters two different tokens, designated as [CLS] and [SEP] tokens where the [CLS] index indicates the start of each and every sequence and the [SEP] token or index serves to separate every pattern in the input. The put in depiction of BERT is the mixture of three types of embeddings i.e. Token, Segment, and Position.

Furthermore, the BERT design, shown in Figure 3.12, trains the language model employing two tasks i.e. Masked Language Model (MLM) and Next Sentence Prediction (NSP). 15% of the tokens in the MLM task are enclosed using [MASK] tokens before the word sequences are passed into BERT. The system subsequently estimates the true worth of the enclosed token. Concerning the task of NSP, the BERT technique accepts input matching set of words and predicts whether or not the secondary sentence in the match corresponds to the sentence following it in the initial sentence. The concept of using a pre-trained model in comparison to traditional methods saves a lot of time for training [139].

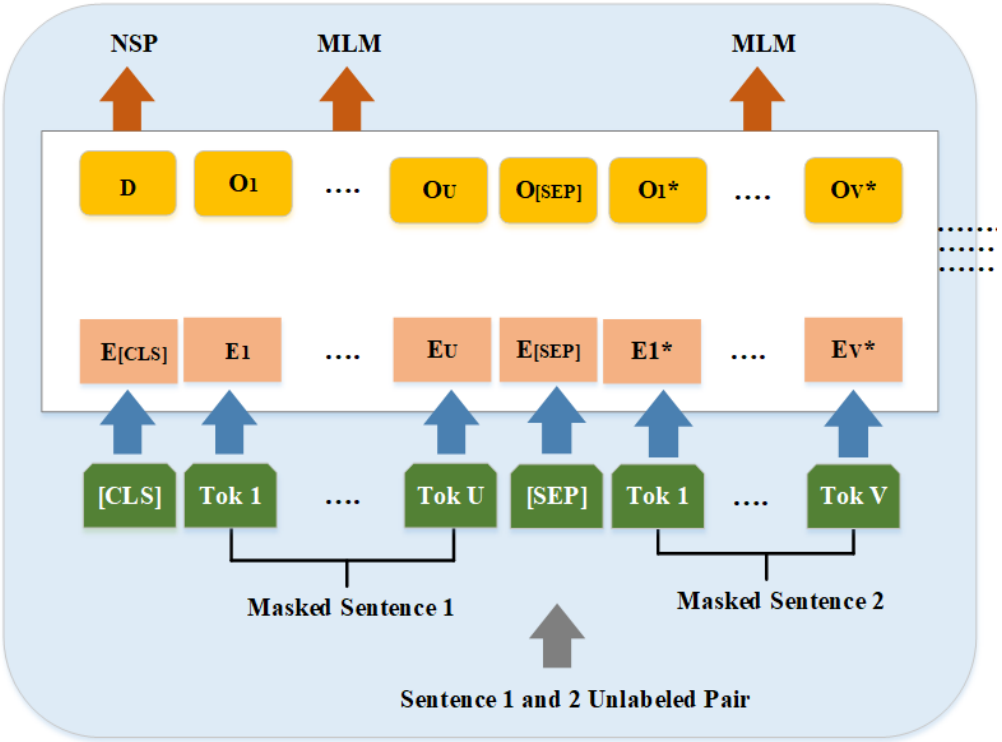


Figure 3.12: BERT Pre-training Architecture

3.3.1.2.1.2 RNN Model

RNN is the widely preferred DL technique for pattern recognition that can handle input data with varying lengths. These models receive output from the previous layer and feed it as input to the current step with the help of the memory concept as shown in Figure 3.13. A recurrent neuron is the most important processing unit of RNN which is responsible for maintaining a hidden state. As text is in the sequential form, therefore, these methods are well suited for text analysis [140].

These models are specially constructed to tackle the issues with serial data and their inside memory feature makes them robust and very precise in estimating what's coming next. One of the disadvantages of these networks is that they suffer from the problem of vanishing gradient. Thus, in this work, two forms of RNN i.e. LSTM and GRU are utilized. LSTM has the capability to save or delete information as per its priority by employing three types of gates. The first gate i.e. Forget gate is

used to drop all the useless details from memory. The next i.e. input gate is responsible for updating new details in memory and the last i.e. output gate selects relevant details and sends them to the next steps. In contrast to LSTM, GRU comprises two gates namely Reset and Update. The reset gate keeps on checking the amount of information that needs to be forgotten and the Update gate checks the amount of knowledge that requisite to be passed in the entire network.

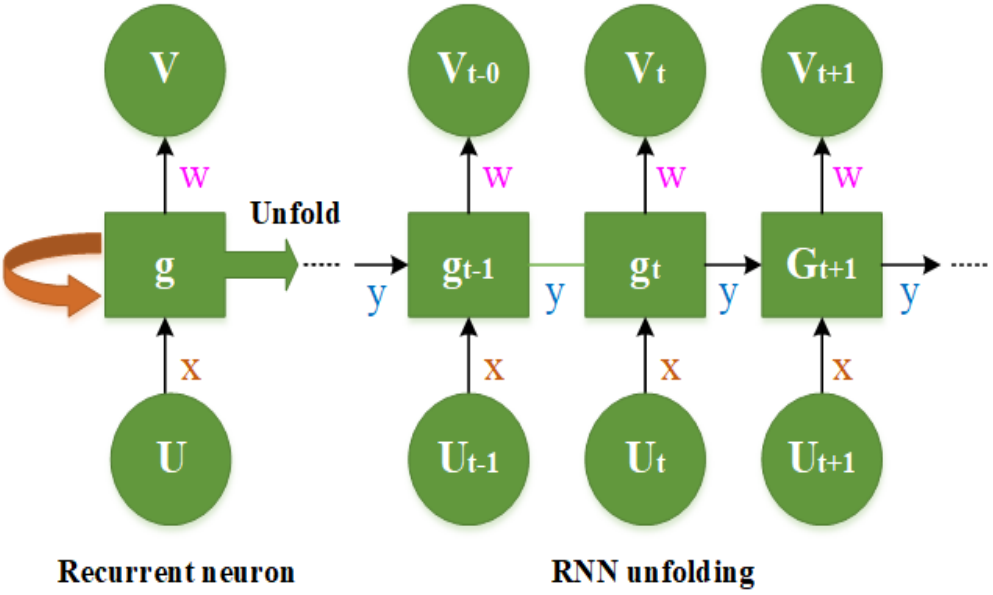


Figure 3.13: RNN Structure

3.3.1.2.1.3 Proposed Model

This section of this chapter proposes to combine BERT and RNN to perform a more reliable classification of depression severity levels. Further, the advantages of LSTM such as accurateness and GRU having low complexity are fused to achieve enhanced performance. A section in Figure 3.11 represents the proposed BERT-RNN framework structure fusing both LSTM and GRU. The process followed is provided in Algorithm 1 depicting the step of the proposed model for depression detection. Firstly, tokenization is performed at the input layer from the input word sequence and the data which is tokenized is then sent to the following layer i.e.

Embedding layer. In this layer, a pre-trained BERT embedding method is employed to generate relevant and useful word embeddings. The output is then transformed into the next or third layer which comprises of fusion of LSTM, Dropout, and GRU layers. Finally, the softmax output layer is used as the fourth layer which provides the likeliness of the considered groups. The category having the greatest probability is considered the forecasted class.

Algorithm 1: Proposed model for depression detection

Input: Text dataset with 3 labels: not-depressed, moderately and severely depressed

Output: Predicted label on test dataset

Step 1. Import the dataset

Step 2. Perform preprocessing of dataset by removing URL's and transforming emoji

Step 3. Divide the preprocessed dataset into training and testing in 70:30 ratio

Step 4. Perform tokenization using BERT

Step 5. Add Embedding layers() with weight

Step 6. Two LSTM() layers

Step 7. Dropout layer

Step 8. LSTM() layer

Step 9. GRU() layer

Step 10. Dropout layer

Step 11. GRU() layer

Step 12. Dense layer() with softmax activation function

Step 13. Train the model on training set

Step 14. Perform testing and evaluation of model on test set

3.3.2 Performance Evaluation

Once the system is created, it is essential to determine its efficacy utilizing some evaluation measures of performance. Depending on the results obtained, the robustness of the applied approach can be successfully analyzed. In this study, measures like accuracy, recall, precision, and F1-score are utilized to evaluate the

efficiency of the proposed approach which are explained in Chapter 2 section 2.6. Accuracy, being, the most preferred measure is considered to make a comparison of the presented approach with the previous studies.

3.3.3 Results and Comparative Analysis

The entire implementation process is performed using Python programming on a GPU system. The dataset is split into 70:30 i.e. training and testing set. The data provided in Table 3.8 are taken as parameters to the BERT with the purpose of embedding extraction. The explained procedure in Figure 3.11 is followed and the results yielded are provided in Table 3. From the data of results, it can be analyzed that the presented hybrid technique attained an excellent performance by achieving an accuracy, recall, precision, and F1-score of 95.0%, 92.0%, 94.0%, and 93.0% respectively.

Table 3.9: Proposed model parameters

Parameter	Value
LSTM layers	3
GRU layers	2
Dropout layers	2
Output function	Softmax
Batch size	256
Optimizer	Adam
Learning rate	0.001
Dropout rate	0.5
Number of encoder layers	4
Max sequence length	512 (typical)
Vocabulary size	30522 tokens
Embedding dimensions	768

To validate results, the proposed model is compared to other models considering solo BERT and RNN which yielded an accuracy rate of 93.41% and 93.49% respectively. Also, BERT is evaluated by combining it with other ML techniques including SVM, decision trees (DT), KNN, and naïve Bayes (NB). The

data in Table 3.9 shows that hybridizing BERT with RNN has given the highest accuracy i.e. 95.38% in comparison to all other techniques applied. The graphical representation of all the outcomes achieved using the proposed and other techniques is depicted in Figure 4. As the BERT model has the benefit of providing improved embeddings, combining it with RNN can yield into improved results.

Table 3.10: Results with different techniques

Model	Accuracy	Recall	Precision	F1-score
BERT	93.41%	91.18%	94.31%	92.43%
RNN	93.49%	92.17%	94.23%	93.29%
BERT-KNN	90.19%	88.28%	91.46%	89.35%
BERT-SVM	92.37%	90.21%	93.39%	91.43%
BERT-DT	88.26%	86.31%	89.41%	87.25%
BERT- NB	86.39%	84.19%	87.14%	85.39%
Proposed	**95.38%**	**92.25%**	**94.43%**	**93.31%**

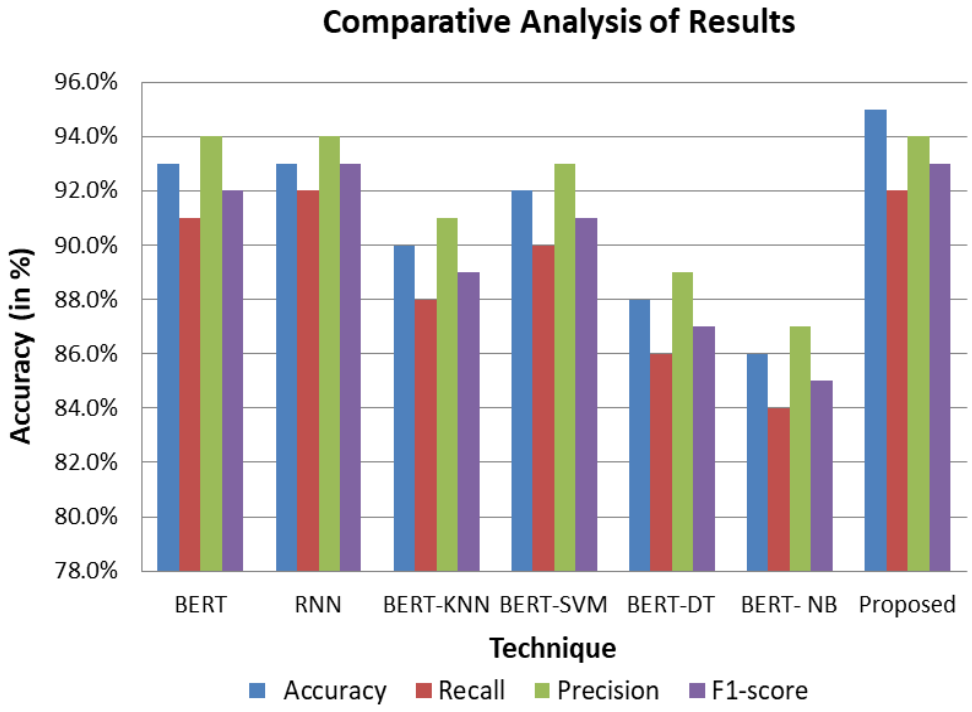


Figure 3.14: Graphical representations of results

Further, the proposed model accuracy is evaluated with the previous studies and the comparative data results are given in Table 3.10 Thus, it is evident from the table below that the proposed BERT-based DL methodology has attained the best accuracy and can be efficiently utilized by clinicians to analyze and treat at different levels of depression.

Table 3.11: Comparison of the proposed approach with previous studies

Reference	**Modality**	**Technique**	**Accuracy**
[54]	Text and audio	CNN	92% (text) 88% (audio)
[129]	Text	BERT, AlBert, DistilBert, RoBerta, Ensemble	61% (Best with Ensemble)
[130]	Text and audio	Logistic regression, SVM	86% (Best with SVM)
[141]	Text	SVM	70%
Proposed	**Text**	**Hybrid BERT-RNN**	**95.38%**

3.4 Chapter Summary

Social media platforms have become the most widely used source for users to share their feelings, emotions, thoughts, views, etc., with their friends and family through pictures, text, audio, videos, etc. Analyzing such data in-depth can provide crucial clues regarding a person's state of mind. Among mental disorders, depression seems to be a rapidly growing disease, particularly in the younger generation

worldwide. Therefore, the early detection of depression is of utmost importance nowadays to save people's lives by providing them with timely counseling and treatment.

This chapter aims to provide two models Firstly, a robust multimodal strategy to detect depression using a hybrid approach for binary classification, In which three models are designed: primarily, BERT for text; then, CNN for images; and third, hybridization of BERT and CNN model for multimodal data, i.e. text + images. The current state-of-the-art analysis revealed very little use of BERT and CNN for text and image data and also attained mediocre performance in depression detection. Therefore, BERT and CNN models were employed individually and combined to achieve higher accuracy. Moreover, a dataset has been created consisting of posts from Instagram to classify depressive and non-depressive users. For text data, BERT and other versions of BERT, namely RoBERTa, DistilBERT, and XLNet are applied and the experimental results show the best accuracy with BERT, i.e., 97.31%. Similarly, for image data, CNN and ML models such as SVM, KNN, and DT are employed and among all, CNN has achieved the best accuracy with 89.42%, thus proving the efficacy of the deep learning CNN model in this research. A robust model that can work on both text and image data is proposed by combining BERT and CNN. The proposed hybrid approach is compared with other combinations, including BERT-SVM, BERT-KNN, and BERT-DT. The results showed that the proposed hybrid approach BER-CNN has achieved the highest accuracy rate, i.e., 99.31%. Lastly, based on the accuracy parameter, it is found that the proposed approach has outperformed the current state-of-the-art studies.

Secondly, to deal with depression, it is necessary to analyze it at different severity levels. So, the later section of this chapter presented a robust BERT-RNN-based hybrid approach to experiment on three labels i.e. not- depressed, moderately, and severely depressed. The text set of data is used which is taken from an online repository and is initially prepared to refine its quality. For classification, the embeddings are extracted using a BERT-based model which is fed into the combination of LSTM, GRU, and a combination of LSTM, GRU, and dropout layers of RNN. The analysis of results obtained using metrics like accuracy revealed the

highest values i.e. 95.38% with the proposed approach in comparison to other techniques as well as state-of-the-art. Thus, the presented strategies in this will be helpful for researchers and doctors to deal with the serious issue of depression accurately. Thus, both the models are aligned with the research objective 2 and bridges the gap of creating the new dataset for the multimodal depression detection and working on severity levels of the depression and now the next chapter will focus on creating a new dataset for Hindi/Hinglish (regional language) data and developing the model for the same.

CHAPTER 4

A MULTIMODAL DEEP LEARNING-BASED FRAMEWORK FOR DETECTING DEPRESSION USING PURE HINDI AND HINGLISH (CODE-MIXED) SOCIAL MEDIA CONTENT

4.1 Introduction

Nowadays, the Internet has become one of the most popular platforms which have revolutionized almost every aspect of life to a great extent. Information on social networking sites spreads so fast and reaches every person within a very short duration of time. Therefore, social media has set new records for being used as the most reliable communication method among individuals. Various online social networking websites including Twitter, Instagram, Facebook, etc. are gaining huge attention from users of every age group to express their thoughts. These platforms are being preferred by people to perform varying tasks such as online shopping, remote education, sharing views and experiences, etc. [97, 142]. Though the use of online sites has taken the world to an advanced level, its severe and dangerous impact can't be ignored. According to a study performed by AlSagri & Ykhlef [104], it is observed that depression and mental case rates have increased in users with high usage of online social platforms. These mental health issues can include stress, anxiety, and unstable thoughts and may give rise to serious actions such as suicide. Thus, it is very necessary to diagnose and cure depression at the initial stages; otherwise, it can be a major issue in ending life.

A study performed by the World Health Organization (WHO) [89] indicated that just a few percent of the 56 and 36 million Indians suffering from depression and anxiety issues can get effective treatments. Most cases remain undiscovered because of the various misbeliefs related to mind health in society. The cause of depression can be anything, but it is analyzed that teenagers are coming into its impact more frequently. Another research released by WHO in 2019 [143] found

that out of a billion people having mental disorders, the world's youth share is 14% of it. The presence of depression is gradually decreasing the people's quality of life and more serious impacts are expected to be arising.

Many depressed people who can't feel comfortable sharing their feelings with friends and family often take the support of social media to reflect their emotional behavior. Social media made it possible to locate such users and the analysis of their posts needs to be evaluated critically to save a person's life. Clinical evaluations for depression using interviews, rating scales, questionaries, and other long subjective procedures can be inaccurate and imprecise, which may result in unwanted delay and faulty analysis [5]. Therefore, automated approaches are required that can perform depression diagnosis efficiently without more overhead and within less time. To automate the process of depression detection, the Machine learning (ML) paradigm is being used by various researchers and has the remarkable capability to perform the classification of users into normal and depressive users utilizing various algorithms [41]. The advancements in Artificial Intelligence (AI), like the evolution of Deep Learning (DL) and Transfer Learning can learn large and complex data within a few seconds. The use of Transfer Learning along with DL can speed up the process of model training on a new task and can result in more accurate predictive analysis [106]. Further, to improve the efficacy of various predictive models, multiple ML and DL models can be combined using different approaches known as ensemble learning. Despite just relying on an output produced by a single model, it can be verified using an ensemble of classifiers to achieve enhanced diagnosed accuracy. So, in a nutshell, due to the tremendous benefits offered by these automated quantitative methods and models, their adoption for depression detection and other unstable mental states can yield remarkable results.

In this area of depression detection, Singh et al. [65] proposed a study to detect emotions in mixed data including Hindi and English language text using Natural Language Processing techniques. The capability of their techniques was evaluated on distinct datasets and they achieved satisfactory accuracy results. In another study by done by Khan et al. [55], a multi-class Urdu database is presented to perform sentiment

analysis considering 9312 reviews. They trained two datasets using different techniques such as word embeddings, rule-based, ML, and DL models, and the results showed the efficacy of pre-trained word embeddings. Chopra et al. [66] proposed a framework to detect hate speech from Devnagri Hinglish language text. They used a Tabnet model-based classifier, which showed its efficacy in the automated classification of mixed code text. Biradar et al. [57] utilized a Deep Network based on Neuron (DNN) for the identification of hate speech. They used libraries to perform transliteration from English to Hindi language, and the proposed architecture achieved the best results.

4.1.1 Major Contributions

A thorough scrutinization of the literature, which is as discussed in the chapter 2, demonstrates some serious gaps that need to be covered to effectively classify depressive and non-depressive posts. Most of the related research work was limited to using pure English language text only. However, it is analyzed that some users prefer only the Hindi language to express their thoughts. Due to the unavailability of a public Hindi dataset and the lack of evaluation of image data along with text data i.e. multimodal data, this study tried to overcome such issues. Further, optimization techniques with DL models and a combination of DL and Transfer Learning have not been explored in the earlier studies for depression detection. Therefore, this chapter aims to present a framework that can efficiently detect depression based on pure Hindi as well as Hindi-English (Hinglish) language mixed data with high accuracy. Some of the main contributions of this study are as follows:

1. To the best of our knowledge, there is no public Hindi dataset exists for depression detection. Therefore, a novel multi-class depression evaluation dataset is proposed and created for the Hindi language, which contains pure Hindi and Hinglish code mixed data including text as well as images gathered from various social media platforms.

2. An approach (CPSO) is proposed that uses a Convolutional Neural Network (CNN) with optimized hyperparameters using a nature-inspired algorithm, namely Particle Swarm Optimization (PSO) for computationally effective classification of image data into depressive and non-depressive posts.

3. A hybrid of transformer-based methods, namely Bidirectional Encoder Representations from Transformers (BERT) and CPSO i.e. BERT-CPSO (BTCPSO) is developed for the identification of depressive and non-depressive posts using Hindi and Hinglish language based on both text and image data.

4. The efficacy of the employed models i.e., CPSO and BTCPSO is compared with other DL, ML, and transformer-based techniques for image and text classification using various evaluation metrics.

5. A comparative analysis of proposed models is also performed with state-of-the-art studies and the results showed the superiority of the developed techniques in depression detection on multimodal Hindi as well as Hinglish language data.

4.2 Proposed Framework

To detect whether an individual is having depression using his posts in the form of text as well as images in Hindi and Hinglish language, a novel methodology is presented in this chapter. The proposed methodology uses hybrid techniques to improve the prediction accuracy and comprises different stages. The entire process of depression detection initiates with the creation of a new Hindi language-based dataset followed by processing of the collected data. Then, the text and image parts are classified separately using different techniques and hybrid approaches. Finally, the efficiency of the presented methodology is checked with other classifiers and models applied in past studies. Figure 4.1 depicts the proposed methodology applied for depression detection and a detailed explanation is provided in sub-sections as follows:

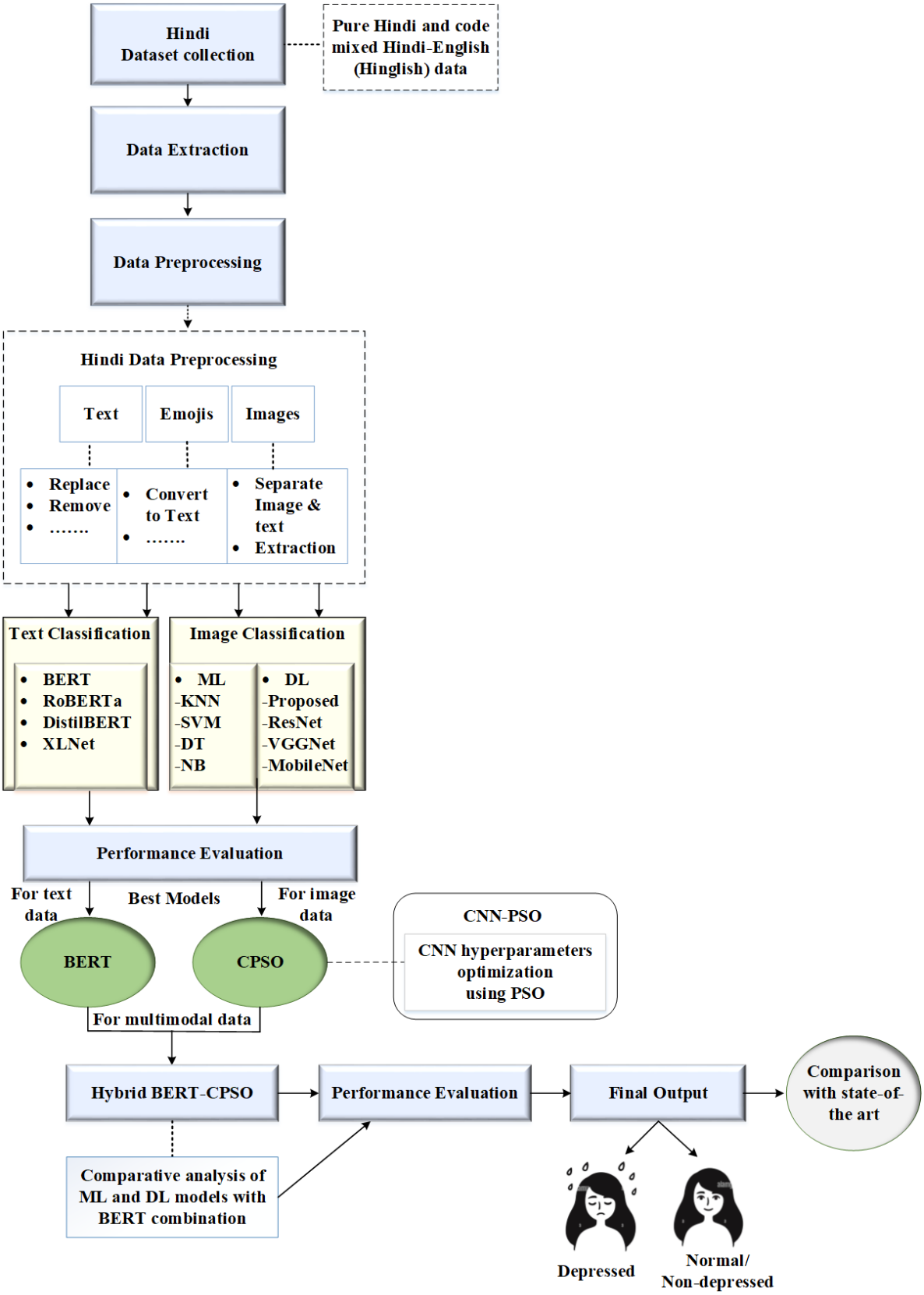


Figure 4.1: Proposed methodology to analyze depression users

4.2.1 Dataset

The robustness of any methodology heavily relies on the dataset's goodness. This research intends to perform the detection of depression using posts of users in both Hindi and Hinglish language, but the analysis of the literature revealed no public existence of such a dataset. Therefore, we created a new dataset based on Pure Hindi and Hindi-English (Hinglish) language, which is one of the main contributions of this research. To accomplish Hindi data collection, different online networking sources such as Instagram, Twitter, Reddit, and LinkedIn were crawled comprehensively. The data uploaded by users, in text, emoticons, and images, either depressive or non-depressive, is gathered to enable multimodal data evaluation. The exploration of collected data showed that along with Hindi, some of the users also post in mixed Hindi-English i.e., Hinglish language. Therefore, this chapter focused on both i.e. pure Hindi and Hinglish multimodal data to diagnose depression efficiently. In all, 11,199 posts were collected for investigation, with 4493 posts comprising only text and 6706 posts containing text as well as images. In addition, out of 4493 purely text posts, 2200 were depressive and 2293 were non-depressive. Similarly, out of a total of 6706 images and text posts, 3565 were depressive and 3141 were non-depressive. The distribution of posts into depressive and non-depressive is shown in Figure 4.2. Table 4.1 presents an overview of samples for the newly created Hindi and Hinglish datasets evaluated in this chapter for efficient depression analysis.

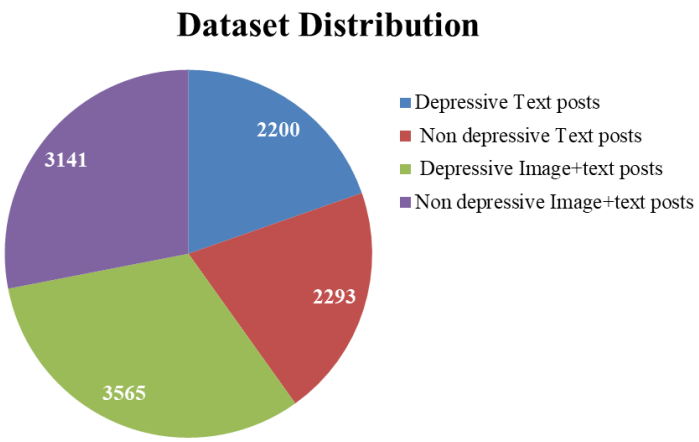
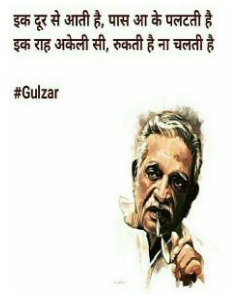


Figure 4.2: Posts distribution of the created dataset into depressive and non-depressive

Table 4.1: Dataset Description

Modality	Sample Examples	Language	Category
Text and Emoticons	मैंने सिर्फ अपने माता -पिता को अपने अवसाद के बारे में बताया और जनरल एक्स लोगों को यह समझने के लिए इतना कठिन है कि यह कुछ ऐसा नहीं है जिसे मैं हर समय नियंत्रित कर सकता हूं या बस टहलने के साथ या अपने दिमाग को व्यस्त रख सकता हूं	Hindi	Depressive
	@Worldofoutlaws मुझे जल्द ही डिप्रेशन मेड्स की जरूरत है, ये रेनआउट मेरे संतुलन को बाहर निकाल रहे हैं	Hinglish	Depressive
	मुझे लगता है कि मुझे पता होगा कि अवसाद क्या है।और शायद ही कभी मैं एक लकवाग्रस्त अवस्था में फिसल जाता हूं, जहां मैं किसी और से दूर एक अंधेरे कमरे में रहना चाहता हूं। लेकिन हर कुछ हफ्तों में, मैं सुबह उठता और कुछ भी सही नहीं लगता।	Hindi	Depressive
	मुझे बुरे दिनों से नफरत है .. जब से मैंने यह कम महसूस किया है .. #Depression #BorderlinePersonality #mentalillness #Bipolar #Anxiety	Hinglish	Depressive
	क्या कोई मेरे अवसाद के मस्तिष्क को ले सकता है और मुझे एक नया एक thx दे सकता है	Hinglish	Depressive
	कोई यह नहीं समझता है कि उसने मेरी कितनी मदद की है।मैं गंभीरता से इतनी बेहतर जगह पर नहीं रहा, मुझे लंबे समय में मानसिक स्वास्थ्य नहीं देखना पड़ा क्योंकि वह मेरे सभी तनाव और अवसाद से राहत देता है।कभी -कभी आपको चीजों को बेहतर बनाने में मदद करने के लिए किसी की आवश्यकता होती है ... इसलिए आभारी ❤	Hindi	Depressive
	खैर क्या एक बिल्कुल बकवास दिन है।मेरे जीवन की बीमार हर कमबख्त दिन मुझ पर शॉट्स ले रही है।मेरी कोशिश करने की बात यह है कि प्रयास करने के लिए मुझे अवसाद, क्रोध और तनाव में गहराई से नहीं खोदना है!ईमानदारी से नहीं पता कि मैं कब या अगर वापस आऊंगा ..	Hindi	Depressive
	एक बकवास रात के बाद खुश होने के लिए उत्सुक था, लेकिन नहीं। कोई बात नहीं।	Hindi	Non-depressive
	मेरी बच्ची को देखने की उम्मीद है	Hindi	Non-depressive
	मैं नहीं चाहता कि मेरी बहन मुझे पूरी गर्मी के लिए छोड़ दे!	Hindi	Non-depressive
	बेटी के प्यारे डायबिटिक बौने हम्सटर की आज सुबह उसके हाथों में मृत्यु हो गई ... सभी जीव प्यार के योग्य हैं।	Hindi	Non-depressive

Images	 	Hinglish	Depressive
Images	  	Hindi	Depressive

			
		—	Depressive
		—	Non-depressive
	 	Hindi	Non-depressive

4.2.2 Pre-processing and Features Extraction

After collecting a sufficient amount of data, the next step is to pre-process the entire data to convert it into relevant and meaningful form. As the accuracy of any methodology heavily depends on the quality of data; therefore, we gave proper care while pre-processing the data by applying various data enhancement techniques. The data gathered had URLs, stop words, spaces, hyperlinks, capital letters, etc. Thus, to upgrade the quality of raw data, we applied various pre-processing methods to process

text, emoticons, and images separately, as discussed below:

4.2.2.1 Pre-processing of Textual data

When it comes to social media data analytics tasks, text pre-processing is typically considered an essential step since it makes text easier to understand with fewer errors, discrepancies, noise, etc., making it less complicated for ML algorithms to process it. Therefore, in this chapter, the pre-processing work begins with improving the text data. The gathered data of 4493 text posts was pre-processed using the way as discussed in the following points [64][144][145]:

1) **Informal Transliteration:** The scarcity of transliteration standards shows the user's difficulty in deciding vowels and other sounds. Further, this research collected a dataset that primarily contains Hindi as well as Hinglish text. Therefore, for the entire Hindi text, transliteration is performed to convert it into English text. For example, the word in Hindi, i.e., 'दस्तावेज़' is transliterated to English as 'dastaavez'. A similar process is followed for all the text written in Hindi to convert it into English text and make it more readable.

   Stopwords are the most frequently used words in the Stoplist that need to be refined or filtered as they don't have much significance. Since Stopwords are not of much importance, therefore, they are removed utilizing the standard list of Stopwords.

2) **Removal of ambiguous and incomplete posts:** All the posts and words that had repetitive meanings and were not complete were then removed from the dataset to make the data more reliable. Further, the abbreviations that may misguide the exact meaning of words were also removed.

3) **Removing URLs and extra spaces:** The links and URLs that do not have any major contribution to the classification process and only add noise to the text were discarded. Also, extra and irrelevant white spaces were removed from the data.

4) **Discarding Special characters, numbers, and punctuations:** The presence of special characters in the text can highly impact the classification results of a model as they are noisy. Therefore, they were eliminated from the text data. Numerical data having less significance was also removed to obtain pure text characters. In addition, punctuations were also discarded because they can be a source of ambiguity.

5) **Conversion to Lower Case:** To avoid any sort of inaccuracy, the concept is employed to transform the inputted text into an identical casing format. Therefore, the entire text was converted into lower case.

4.2.2.2 Pre-processing of Emoticons and Emojis

Similar to text, emoticons and Emojis posted by users on social media websites have great significance in differentiating a depressed person from a normal person. An emoticon is frequently made up of a collection of punctuation, letters, and numbers that have been structured to resemble an individual's face. As an example, the symbol :-O or 😲 signifies surprise. A pictogram that could represent anything constitutes what an emoji is. Emojis and emoticons were then replaced in place of descriptive text to obtain appropriate data about user emotions. For example, the emoticon 🖤 is converted into text and is written as 'black heart'. Once the emoticons are transformed into the text, now, these are appended with the tokenized text due to their similar format.

4.2.2.3 Pre-processing of Image data

To enhance the efficiency and robustness of a diagnostic model, the combination of different modalities can be utilized. Images can directly reflect the mental state of an individual, just by visualizing them. Therefore, in this chapter, image data was also considered and pre-processed separately, along with text and emoticons. The analysis of social media posts indicated that users also use text on the image data

to express their emotions and sentiments. Thus, it is necessary to separately extract text from the image so that it can be processed efficiently.

In this chapter, an "Optical Character Recognition" (OCR) [111] approach is applied for text and image separation and extraction. This technique scans the text from image files and provides the required text file to use it in further analysis. The used OCR system can take image files of any dimension as input; however, attention should be given because fuzzy or deformed images might not give optimal results. The block diagram showing the OCR process used for extracting text from the given images is as shown in Figure 4.3. The textual files obtained from image data are then further processed using similar steps, as employed for text pre-processing. For the image part, to perform analysis and processing, several ML and DL approaches are applied. The hybrid optimization with the DL technique is also proposed to evaluate images with high accuracy.

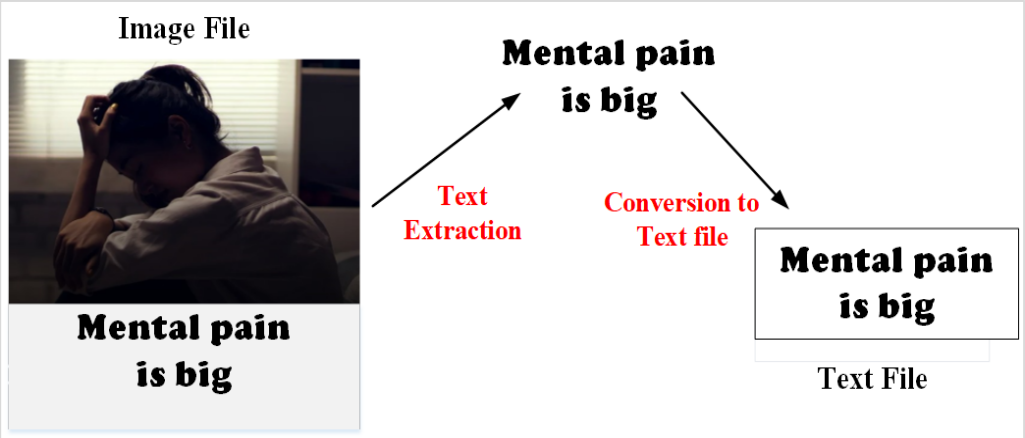


Figure 4.3: Block diagram of OCR for text extraction from image

After pre-processing the entire Hindi and Hinglish data in text, emoticons, emoji, and images; the next step was to extract the informative parameters or features from all modalities to classify depressive and non-depressive posts. The features from text data are extracted by employing various versions of BERT and other approaches, while a hybrid CNN-PSO (CPSO) technique is proposed to extract optimal image

features. Lastly, a combination of text and image models (BERT-CPSO) is presented that can correctly classify multimodal posts into normal and depressed based on Hindi and Hinglish data. A thorough presentation of the techniques proposed and employed is provided in the next sub-section.

4.2.3 Classification

In AI and ML, classification is a critical issue that entails classifying data items depending on their properties or attributes into specified groupings or subcategories. Developing a model that can correctly assign categories and labels to new, unexpected data items is the primary aim of classification. Depending on the particular issue and dataset, identifying an appropriate classifier and evaluation metrics often requires testing and alterations to achieve the best results. In this chapter, different models are proposed and employed to classify text and image data separately. This sub-section gives an explanation of the proposed techniques used for classification purposes to perform accurate categorization of depressive and non-depressive posts.

4.2.3.1 Classification of Textual Data

The text-based posts on social media that have been obtained are analyzed employing a variety of TL strategies, like BERT, RoBERTa, DistilBERT, and XLNet. In the past few years, Transfer Learning has demonstrated significant improvements in the categorization of texts and other computer vision tasks. A classifier that has been evaluated and trained for a particular job is subsequently altered for a similar task using the ML approach known as transfer learning. It takes advantage of the representations and knowledge acquired in a specific field to improve performance in another [113]. Therefore, to find the model that produces the best results, each adopted model, as described below, is trained and tested individually on textual data.

1) **BERT and its variants:** Researchers have come up with several kinds of approaches

for training general-purpose linguistic representation systems using the enormous amount of unannotated text on the web to fill up the data gap. A powerful pre-trained computational language model termed BERT (which stands for Bidirectional Encoder Representations from Transformers) has been successful in several NLP problems, including text categorization. BERT is an increasingly common choice for the classification of text tasks due to its capacity to derive contextual and semantic from text. The transformer architecture, which has revolutionized NLP, is the cornerstone of BERT. The transformer design has become known for its efficacy in extracting context from input sequences [114]. The fact that BERT is bidirectional is one of its unique characteristics. It considers the context of every word in a phrase from the right as well as the left side. BERT operates superbly on a variety of NLP tasks due to this bidirectional contextual knowledge. To get excellent results in NLP problems, the BERT method for transformers frequently and drastically reduces the amount of labeled data and training time required.

Unsupervised pre-training and supervised task-specific fine-tuning constitute the two phases of BERT. BERT is trained unaided on a significant amount of text data throughout the pre-training phase. By predicting words that are missing (masked language modeling) and interpreting the connections between phrases (next sentence prediction), BERT gains comprehension of the architecture and semantics of actual language at this phase. BERT takes out an extensive number of characteristics during pre-training that are employed to encode both contextual data and semantic meaning. Phrases and words are represented by the model as extremely dimensional vectors all over time. BERT frequently analyses the newly received text in layers, such as twelve layers for BERT-base and twenty-four for BERT-large [146]. Taking into consideration of the context from previous stages, each layer enhances the representations of the tokens. The BERT-base version, which has 12 layers of encoders layered on top of one another and 110 million parameters, is utilized in this study. Using labeled information, BERT could be optimized for certain downstream tasks in natural language processing after pre-training on an immense quantity of text data. The already trained BERT model is fine-tuned by incorporating task-specific layers and training it on the desired job. The algorithm's ultimate result is the class label, which

corresponds to depressive and non-depressive posts and can have a value of either 0 or 1.

In a nutshell, BERT's strength dwells in its ability to extract deep contextual and semantic information from substantial pre-training. With very little task-specific input from the user, it learns to represent concepts and words in a large vector space that may be adjusted to various NLP applications. This method has significantly improved NLP tasks and has become an essential part of current NLP models [147]. The pattern of information flow for a word in the BERT approach is depicted in following Figure 4.4, where the embedding representation is ED1, the final result is F1, and the intermediate representations of the same token are Trm. Some of the differences among these techniques employed for text classification are summarized in Table 4.2.

2) **Robustly Optimized BERT Pre-training Approach (RoBERTa) model**, intended to overcome multiple challenges and boost BERT performance. RoBERTa is renowned for its robustness. RoBERTa is pre-trained on a much bigger collection of text data

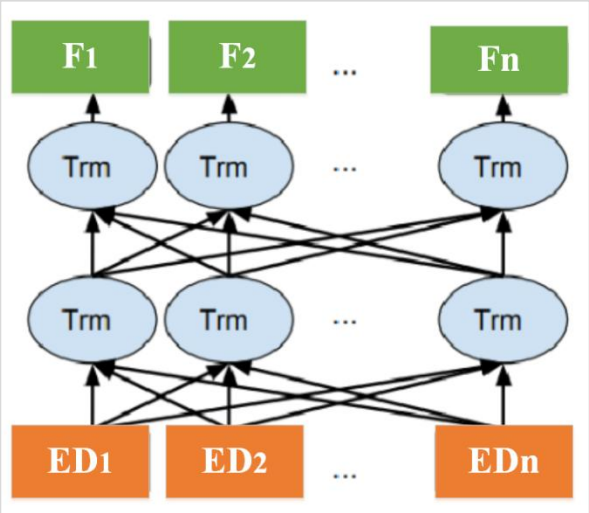


Figure 4.4: Flow of information in the BERT model to evaluate text [114]

compared to BERT. To create more robust language representations, it employs a wide

range of text sources, including BookCorpus, and other freely accessible material. Furthermore, unlike BERT, RoBERTa uses adaptive masking during pre-training instead of a predetermined masking pattern. This indicates that RoBERTa chooses multiple masks at random for each batch of training data, providing a more diverse learning experience. During pre-training, RoBERTa removes the Subsequent Sentence Prediction (SSP) task. The omission of Next Sentence Prediction (NSP) has been proven to be advantageous in the training process. In contrast with BERT, RoBERTa leverages a bigger batch size as well as a learning rate. These hyperparameter tweaks enable RoBERTa to reach convergence faster and perform better. RoBERTa maintains compatibility with the BERT design, making subsequent tasks relatively easy to fine-tune using existing BERT-based programs and techniques.

Table 4.2: Differences among BERT and its variants [147]

Comparison factors	**Models**			
	BERT	**RoBERTa**	**DistilBERT**	**XLNet**
Size (million)	110M	125M	66M	110M
Embedding layer size	30,522*768	50,265*768	30,522*768	32,000*768
Training time	Base: 8*V100*12 days	Large:1024*V100* 1 day	Base:8*V100*3. 5 days	Large:512 TPU chips *2.5 days
Pre-trained data	16 GB BERT data (BookCorpus + English Wikipedia), 3.3 billion words	160 GB (16 GB BERT data + 144 GB other)	16 GB BERT data, 3.3 billion words	16 GB BERT data
Technique	BERT with Masked Language Model (MLM) and NSP	BERT without NSP	BERT Distillation	BERT with permutation -based modeling

3) **DistilBERT** is a resource-effective and distilled form of BERT that was created to make advanced NLP capabilities accessible in situations where computing resources

are constrained. DistilBERT decreases model size through a process referred to as knowledge distillation. It was specially trained to replicate the behavior of the larger BERT model while possessing a smaller number of parameters. DistilBERT is more quickly to train, deploy, and execute inference because of fewer parameters. In comparison with BERT, DistilBERT has a simplified architecture and fewer transformer layers, minimizing the model's size and computational requirements. DistilBERT approximates a more extensive neural network with a smaller version using the distillation process [116].

4) **XLNet** is a transformer that brings together the positive aspects of autoregressive modeling and auto-encoding to compensate for their drawbacks. Rather than implementing a rigid forward or backward sequence for factorization as in conventional models using autoregressive idea, XLNet optimizes the sequence expected log probability, incorporating every possible order permutation of factorization. It is a refined NLP model that integrates permutation-based training into its transformer structure. It specializes in determining connections between words in a phrase and capturing bidirectional context. XLNet, on the other hand, can evaluate context from any point in a phrase, unlike BERT, which relies on a predetermined masking technique. This improves XLNet's ability to comprehend the associations between words in a phrase [117]. XLNet combines segment recurrence method of Transformer-XL's and proportional approach of encoding into pre-training, which boosts performance, particularly for tasks involving an extended text sequence.

4.2.3.2 Classification of Image Data

An accurate analysis of images may provide clear indications of the presence of an aberrant mental state. Numerous algorithms based on ML and DL platforms have been developed for analyzing image data. However, ML techniques have demonstrated notable effectiveness in differentiating normal and depressive posts, but they additionally need feature extraction and domain expertise. Deep Learning, in contrast, reveals the ability to train on real-time data without the explicit requirement of feature extraction. Therefore, due to the tremendous capabilities of DL-

based architectures, this chapter proposed a hybrid CNN-PSO-based technique (CPSO) to perform efficient classification of gathered image data. The hyper-parameters of the employed CNN model are optimized using a swarm intelligence approach known as Particle swarm optimization (PSO) to produce a more robust analysis. Further, the results obtained using the proposed CPSO technique are also compared with other DL and ML algorithms to find the best one. The description of the proposed approach and each technique applied is comprehensively provided in this sub-section.

(i) Convolutional Neural Network (CNN)

The CNN design is one of the most outstanding instances of an Artificial Neural Network (ANN), which is frequently employed to tackle complex image-based pattern recognition challenges. It can learn spatial arrangements between components, including edges, textures, and forms that are crucial for identifying objects in images. In this study, CNN is chosen above other models since it allows for the change of its hyperparameters, which can increase accuracy to a greater extent. CNNs process the data that comes in utilizing several connected layers, which are stacked on one another [118][119]. The input, convolution, non-linearity, pooling, flattening, and classification or fully connected layers make up the basic CNN architecture as shown in following Figure 4.5.

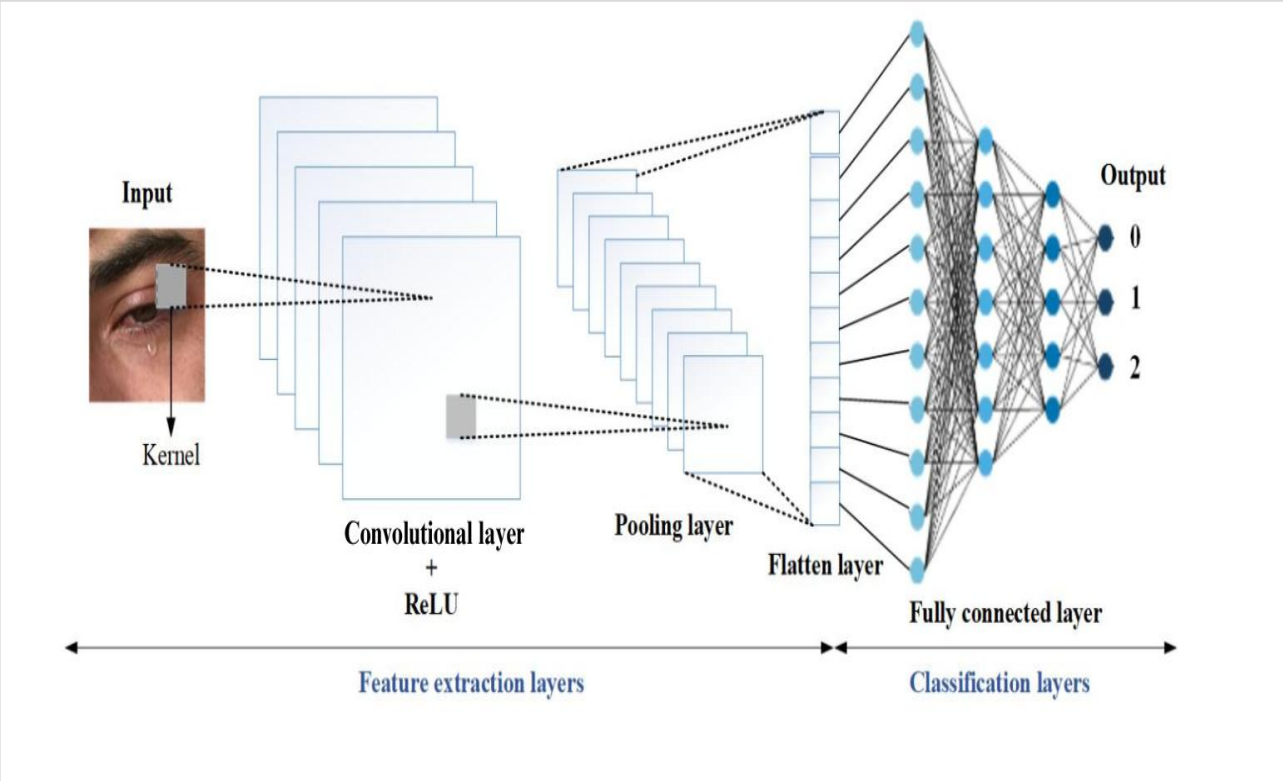


Figure 4.5: Basic architecture of CNN model [118]

1) Input Layer

The input layer typically is CNN's first layer, where videos or images that the neural network will process to extract its characteristics are inputted. Two-dimensional matrices are utilized for keeping all information. This layer reflects the pixel matrix of the image and accepts the input information from the external world.

2) Convolutional Layer

Convolutional process is a mathematical operation that takes place at the convolutional layer between the input image and a filter of a particular dimension, such as MxM. It imparts to the kernel essential characteristics such as the ability to recognize lines, edges, focus, blur, curves, and colors, among others. The activation function in the convolutional layer has an identical idea to the activation utilized in

any neural network. Different activation functions exist; one of the most prominent in these types of designs is the Rectified Linear Unit (ReLU) function which adds non-linearity to the model and allows it to recognize deeper patterns in the data. With I, K, q, and r as the input image, kernel, row, and column of the image, the convolutional procedure is expressed in equation (4.1) [118].

$$C(u,v) = (I * K)(u,v) = \sum_q \sum_r I\,(q,r) K(\text{u-q}, \text{v-r}) \tag{4.1}$$

3) Pooling Layer

The key objective of the Pooling Layer is intended to decrease the size of the convoluted feature map to reduce computational expenses, which is accomplished separately on each feature map and by minimizing the connections between layers. Max pooling, one of the most prevalent grouping techniques, has been used in this study to identify the feature map pixel with the highest value. Assuming we have a 4*4 future map, the pooling operation is usually executed employing a 2*2 filter [119].

4) Flatten Layer

The produced 2-dimensional array from pooled feature maps is all smoothed into a single, continuous linear vector with the help of a flattened layer. To categorize the image, the flattened matrix serves as input to the fully connected layer.

5) Fully connected or Classification Layer

These layers comprise the final couple of stages in CNN architecture and frequently appear before the output layer. The flattened vector persists via a few more FC levels, where the standard operations on mathematical functions occur. Normally, overfitting in a training dataset can occur from all features being connected to the FC layer. To address this overfitting problem, a dropout layer is introduced, in which only a handful of neurons are eliminated from the neural network while training, lowering

the overall size of the model. Fifty percent of the nodes in the neural network are arbitrarily deleted upon passing a dropout of 0.5 [119].

Three convolutional, three max-pooling, a flatten, and two output or dense layers were employed to generate the CNN in this study, which is then followed by a dropout of 0.5. With varying numbers of filters, the first, third, and fifth positions utilize the three convolutional layers. The second, fourth, and sixth places all include three max-pooling layers. In the suggested CNN network, a flattened layer is added at position seven, followed by two completely linked layers at places eight and ten. The provided CNN is based on the CNN model's fundamental structure, which is shown in Figure 4, and it has been modified for this study, as shown in Figure 4.6.

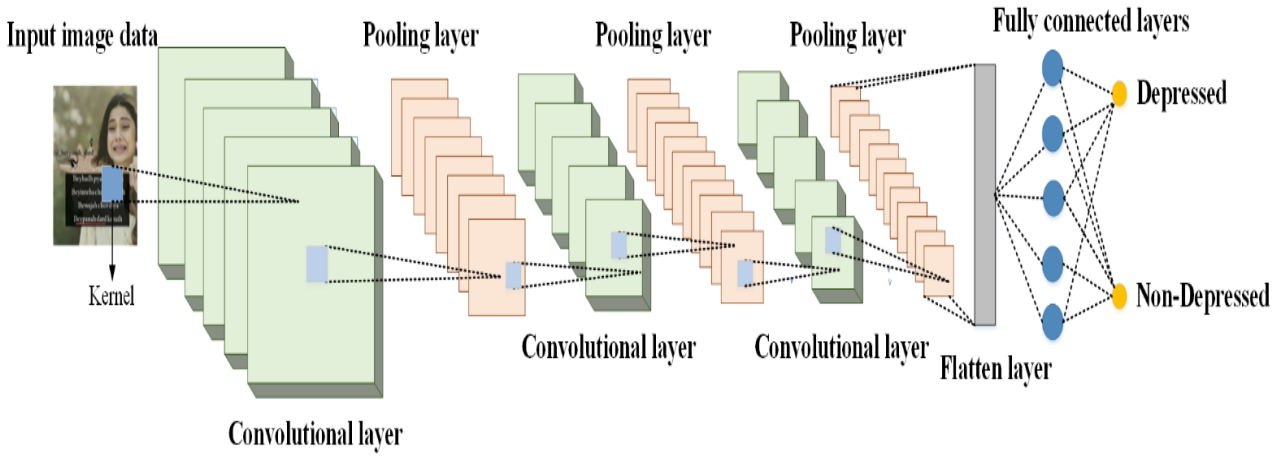


Figure 4.6: Proposed CNN architecture

(ii) Particle Swarm Optimization (PSO)

PSO, established by Kennedy and Eberhart in the year 1995, is a bio-inspired algorithm that looks for the most effective solution in a given space. It varies from other optimization techniques in that it demands just the objective function and doesn't rely on the gradient or any special form of the objective function [148]. Each bird is depicted by particles that "move" in a multidimensional search space and

"adjust" according to their knowledge of neighbours and your own. It serves as a stochastic method based on the collective intelligence of the swarm and is inspired by the way birds look for food. The PSO candidate solution is commonly referred to as a particle and has a relationship with a certain velocity and position. PSO makes use of a fitness function and the particle's velocities to identify the best possible outcome through integrating local and global information. The particle updates its location according to its optimal values, which are stored. Thus, this algorithm functions by storing and sharing the best particles on a local and global scale [149].

In other words, in PSO, every single particle is modified after each iteration using the "best" values. The first choice stores the fitness value and is the finest fitness solution referred to as "pbest". The next "best" value determined by any particle in the population is tracked by the particle swarm optimizer. The term "gbest" denotes this best value as a global best. Fig. 8. illustrates the movement of the particle using the concept of PSO. The following equations are used by the particle to update its position and velocity after selecting the two optimal values.

$$s_i(t+1) = s_i(t) + v_i(t+1) \tag{4.2}$$

$$v_i(t+1) = v_i(t)w + c_1 i_1(y_i - s_i(t)) + c_2 i_2(\hat{y} - s_i(t)) \tag{4.3}$$

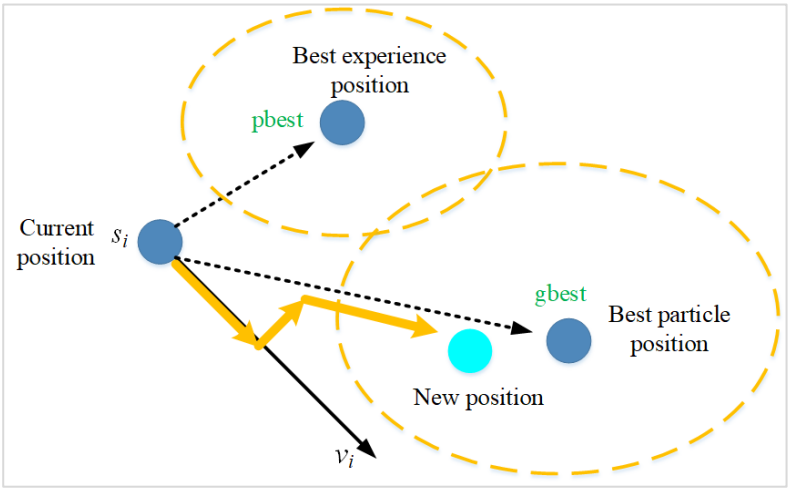


Figure 4.7: Showing the concept of PSO [148]

In equation (4.2), si (t) denotes the particle 'i' position in time 't'. Equation (4.3) represents the velocity update where 'v' indicates the velocity of particle 'i'. Here, c1 and c2 are constant integer values indicating cognitive and social components. i1 and i2 denote random values in the interval [0 1], w is the inertia weight, yi, and y^ represent the local and global best positions of the particle. The pseudocode of the PSO algorithm showing the entire process is given in Pseudocode 4.1.

Pseudocode 4.1: PSO

```
Begin
   For each particle
      Initialize parameters (position and velocity) randomly
   End For
do
   For each particle
      Compute the fitness function value
      if the objective fitness value is better than local best fitness value (pbest) in past, set the
      current fitness value as new pbest
      End if
   End For
   Select the particle with global best fitness value (gbest) from the entire neighborhood
   For each particle
      Update the position s_i of particle using equation (2)
      Update the position v_i of particle using equation (3)
   End For
until termination criteria is reached
End Begin
```

For each particle, the fitness is evaluated and the values of local and global bests are updated. Depending on the values obtained, the position and velocity for each particle are then modified and the procedure is repeated until the stopping criterion is met.

(iii) Proposed CNN-PSO (CPSO) Technique

As stated earlier, one of the significant and unique features of CNN models is the ability to adjust hyperparameters. These models are more efficient at achieving accurate classification since the parameters, such as learning rate, dropout, momentum, number of filters, filter size, batch size, and the number of layers in a CNN, can be simply modified according to the scenario. Therefore, to achieve great performance for identifying depressive and non-depressive posts, this chapter aims to choose optimal CNN parameters by applying PSO. In the present chapter, four hyperparameters of CNN such as learning rate, batch size, number of filters, and number of iterations, are taken to optimize them using PSO. The flowchart depicting the general PSO and the proposed CPSO architecture is presented in following Figure 4.8.

The PSO is initialized in this procedure in accordance with the execution parameter, creating the particles. Each solution demonstrates a full training session for CNN as each particle is a potential solution and each position has a parameter that can be improved and optimized. The training technique is an iterative process that terminates when every single one of the particles produced by the PSO for every generation is analyzed. The computational expense is higher and is affected by the database dimensions, particle size, number of PSO repetitions, and the quantity of particles in every iteration of the process. In simple terms, if the PSO runs with 20 particles and 20 iterations, the CNN training procedure is executed for a total of 400 times.

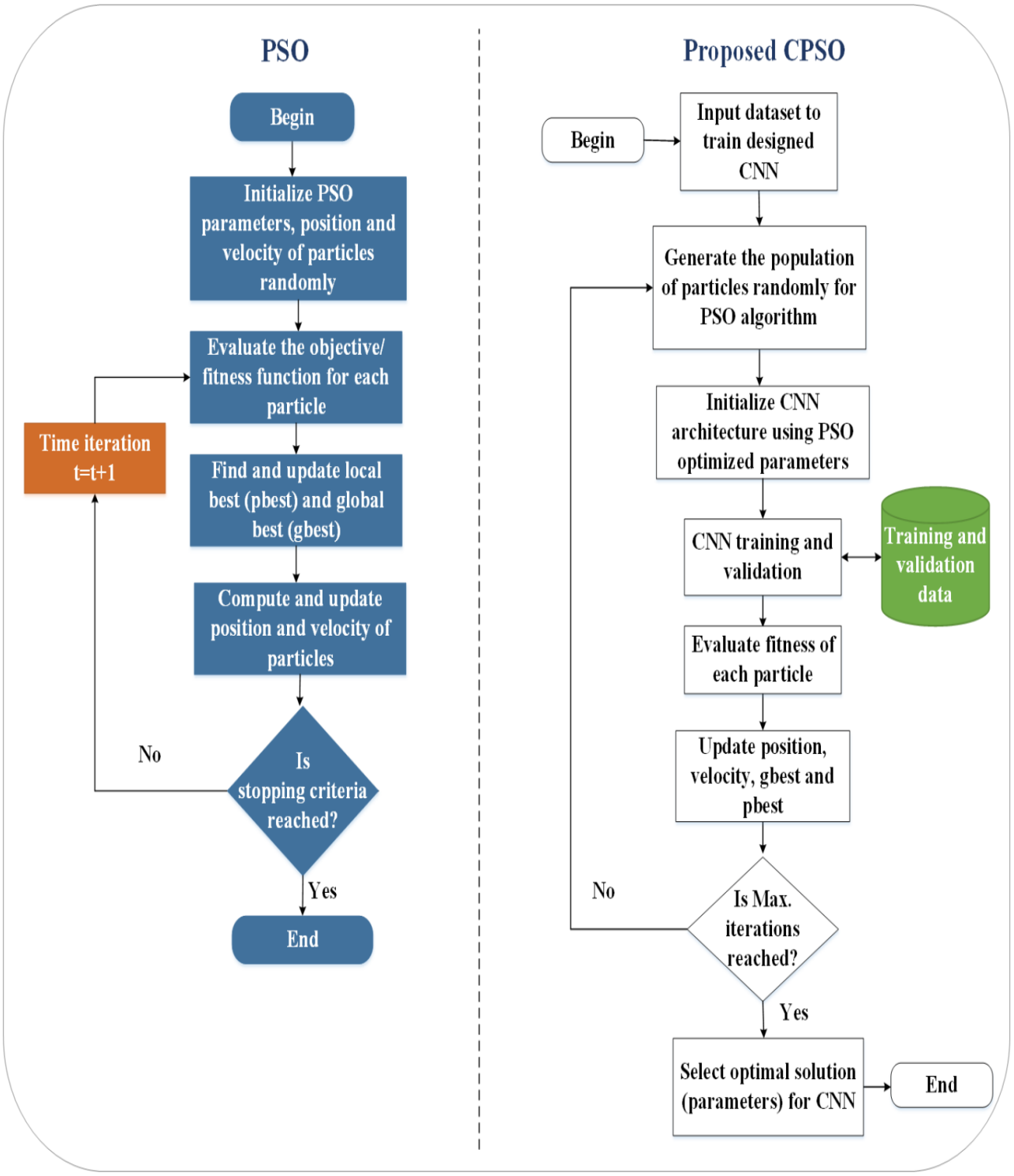


Figure 4.8: Flowcharts of Basic PSO (left) and the Proposed CPSO architecture to optimize CNN parameters using PSO (right)

The steps followed for optimizing the CNN employing the PSO algorithm are provided below:

(1) Load the gathered dataset as the input of a comparable type and properties to train the CNN. This phase comprises importing the previously described pre-processed Hindi and Hinglish datasets and categorizing them for the CNN. In a nutshell, the data associated with the image that is going to be used contains similar characteristics for scale, pixel size, color attributes, and file format.

(2) Establish the particle population randomly for the PSO method; the particle design is part of this stage of the process. The tuning of PSO is performed using different hyperparameter values and the best ones are selected which is mentioned in Table 4.3. The PSO parameters are configured to take into account the total number of iterations, the number of particles, the inertial weight, the cognitive constant (c1), and the social constant (c2). Tuning the PSO on these values helps the algorithm to converge at the faster rate to find the best global solutions.

(3) Configure the architecture of CNN with the optimized PSO parameters i.e. learning rate, batch size, number of filters, and iterations. Initialize and tune the CNN algorithm based on the best hyperparameter values stated in Table 4.3. The number of layers and other parameters such as convolution layers, dropout, activation function etc., are experimented by altering their values and position. Finally, the CNN tuning provided the optimized position of the considered layer blocks and other parameter values; where the model has shown the highest performance.

(4) Training and evaluation for CNN. An accuracy rate for each model is calculated when the CNN reads and analyses the input dataset, using images for training, validation, and testing. As the outcome of the objective function's operation, these values are sent back to the PSO.

(5) Analyze and evaluate the desired objective function (accuracy rate is taken in this case). For finding the optimum value, the PSO algorithm explores the objective function.

(6) PSO parameter updates. Based on its individual most well-known position (Pbest) in search space and the best-known location of the entire swarm (Gbest), every particle modifies both velocity and position at every round of iteration.

(7) Once the stop situation (in this scenario, the maximum number of iterations) is reached, the process begins again, evaluating each particle.

(8) The best solution is finally chosen. The CNN model in this process selects particles that are represented by global best i.e. Gbest.

Table 4.3: Tuned hyperparameters of CNN and PSO

CNN Hyperparameters	**Selected Values**
Convolution Layers	n=3: {1,3,5}
Pooling Layers	n=3: {2,4,6}
Flatten Layers	n=1: {7}
Fully Connected Layers	n=2: {8,10}
Dropout Layer	n=1: {9}
Convolution Layer Filter	{32,64,128}
Dropout	0.5
Learning Algorithm	Adam
Nonlinearity Function	ReLU
Output Layer Activation Function	Sigmoid
Number of Epochs	10
PSO Parameters	
Inertial Weight	0.6
Particles	10
Threshold	1e-6
Number of Iterations	100
Cognitive Constant (c_1)	1.0
Social Constant (c_2)	1.5
Upper Limit	0.1
Lower Limit	0.001

(iv) Comparative ML and DL models

After the implication of the proposed CPSO model, the results are comparatively analyzed by employing other ML and DL techniques to find the best one. For such purpose, different ML algorithms including k-nearest neighbour (KNN), Support vector machine (SVM), Decision trees (DT), and Naïve Bayes (NB) were employed. Similarly, ResNet, VGGNet, and MobileNet DL-based models are adopted to make effective comparisons. An overview of these techniques is provided as follows:

a) Machine Learning (ML) based Techniques

1. SVM: An effective supervised Machine Learning strategy used for classification is called the Support Vector Machine (SVM), which works optimally when there is a noticeable gap of distinction between the categories or when there is no linear manner to divide the data. SVM attempts to determine the optimal hyperplane in a feature space with high dimensions for distinguishing data points from different groups. In other words, a hyperplane (a selection border) that improves the gap between two distinct categories is identified using SVM. This hyperplane is known as the 'support vector machine.' Mathematically, the equation for the SVM using a linear kernel can be given as equation 4.4. [121].

$$f(y) = wy + b_i \tag{4.4}$$

Here, f(y) represents a decision function; y and w are the feature and weight vectors, and bi denotes the bias due to which the hyperplane moves away from the origin.

2. KNN: A simple and intuitive supervised Machine Learning procedure called k-nearest-neighbors (KNN) is being used for classification applications. It is a part of the instance-based or slow-learning algorithm group. Following the similarity of data points in a feature space, KNN generates predictions. The core concept of KNN is the concept that identical data points often fall within the same class or possess similar goal values. The first step in KNN is to pick which neighbors (K) will be considered into account while

producing predictions. Hyperparameters 'K' is often selected through cross-validation. Then, the distance between every point of data in the practice dataset and that data point to be inferred is determined. Utilizing a distance metric (like the Euclidean distance) in the feature space, KNN tracks the K nearest data points (neighbors) in the training dataset for the selected data point [120]. Afterward, KNN assigns the group with the highest frequency to the estimated point of data simply measuring the frequency of every group among the K nearest neighbors referred to as majority voting.

3. NB: Naive Bayes (NB) is an extensively utilized, easy-to-understand probabilistic method for classification in the fields of Machine Learning and data analysis. It relies on Bayes' theorem and is especially suited for image categorization and other qualitative problems with classification goals [150]. The Bayes theorem is a classification theory that can be defined as follows:

$$P(v|u) = (P(u|v) * P(v)) / P(u) \quad (4.5)$$

In equation (4.5), P(v | u), P(v), and P(u) are the posterior, prior, and evidence probabilities of group v given input u, and P(u | v) denotes the likelihood of determining input u. Given the group label, NB suggests the presence or lack of a particular attribute is independent of the presence or lack of other characteristics. The probability computations are significantly simplified as a consequence. NB is inexpensive in terms of computation and appropriate for high-dimensional data.

4. Decision Tree (DT): A flexible and comprehensible machine learning technique used for categorization tasks is decision trees. They present a tree-like visual representation of the decision-making process, with each node representing a decision to make or test on an identifiable attribute and each leaf node giving a class label [122]. Utilizing a splitting criterion, which controls how the data is broken down into subsets, decision trees make selections at every internal node and Gini Index constitutes one of the most popular separation criteria. Complicated boundaries for decisions and irregular relationships in the data are effectively handled by decision trees. When the tree is overly deep, they are susceptible to overfitting, which can result in poor generalization. To prevent this, pruning

techniques are employed to simplify the tree while enhancing its capacity for generalization. Additionally, they provide the foundation for ensemble approaches like Random Forests and Gradient Boosting, which leverage many decision trees to boost the accuracy and dependability of estimates. The hyperparameters utilized for the employed ML models are given in Table 4.4.

Table 4.4: Hyperparameters used for the comparative ML Models

Method	**Hyperparameter**	**Selected Value**
SVM	Complexity parameter, C	10
	Type of Kernel	Linear
	Gamma	Auto
	Scale	Logarithmic
KNN	Distance metric	Euclidean
	No. of Neighbours 'k'	5
NB	Alpha	0.01
	Fit_prior	True
DT	Criterion	Gini
	Feat_max	sqrt
	Depth_max	7

b) Deep Learning (DL) based Techniques

1. ResNet: Targ et al. [151] revealed ResNet in 2015, an abbreviation for "Residual Network," a deep convolutional neural network design, developed to address the obstacles of training in complex neural networks. At the core of the network is the revolutionary idea of bypassing connections or residual links that allow information to travel effortlessly across the network's levels. ResNet effectively addresses the problem of vanishing gradients that hinder the learning of very deep networks through the inclusion of these

links. The remaining blocks, which are composed of an initial path with layers of convolutional layers and a short path, constitute the basic structure of the design. ResNet architectures have significantly raised the expectations for identifying image classification tasks and can be extremely deep with hundreds of layers. There are many distinct versions of ResNet, such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The network's layer number is denoted by the number that appears in the name.

2. VGG Net: The Visual Geometry Group (VGG) at the University of Oxford invented the deep CNN architecture known as the VGG. In the domains of image processing as well as DL, VGGNet is well-known for its simple nature of access and performance. It was initially released in 2014 and features a consistent layer structure with tiny 3x3 convolutional filters and regular 1-pixel padding. With deviations like VGG16 and VGG19 comprising 16 and 19 layers, respectively, VGGNet has been identified for its depth. This depth permits the network to understand intricate traits and patterns, allowing it to be an effective tool for identifying objects, image grouping, and other operations. Transfer learning additionally makes significant use of pre-trained VGGNet models, which facilitates fine-tuning to tackle particular computer vision task [152].

3. MobileNet: Especially for portable and embedded systems, MobileNet is a family of deep CNN design made to facilitate efficient on-device learning algorithms. MobileNet, which was created by Google researchers, seeks to reconcile model accuracy alongside computational efficiency. MobileNet has become renowned for being able to deliver precise estimates while being compact and qui ck. The distinctive feature of depthwise separable convolution, which partitions conventional convolutions into a pair of distinct operations, i.e. depthwise and pointwise convolution, is at the foundation of MobileNet's efficacy. This type of design is ideal for devices with restricted computational and memory capacities as it substantially drops the amount of computation required and the total amount of parameters. By the use of hyperparameters like width multiplier and resolution multiplier, MobileNet offers adaptability and allows customers to adjust model size and computational cost to fulfill their particular needs. With the help of MobileNet, edge devices are capable of finishing tasks like image and object categorization, more

accurately and efficiently [153]. Table 4.5 presents the hyperparameters utilized for the employed DL models.

Table 4.5: Hyperparameters used for the comparative DL Models

Parameter	VGG19	MobileNetV2	ResNet50
Input	(32,32,3)		
Epochs	15		
Classifier	Softmax		
Batch size	128	64	128
Optimizer	SGD	RMSProp	Adamax
Regularization	Batch Normalization	Nil	L2 Regularization
Loss function	Binary cross entropy		

Following the implementation of all text and image analysis approaches, their effectiveness is assessed using evaluation metrics. The goal was to identify the most effective strategy among all the models employed for text and graphics individually. In contrast to other approaches, the results obtained as presented in subsequent section shows the best accuracy with BERT for text data and the proposed CPSO for image data. As a result, a new model is suggested that classifies multimodal data comprising both mixed image and text data by combining these two superior approaches i.e. BERT-CPSO (BTCPSO).

4.2.3.3 Hybrid BTCPSO Technique

Due to their superior accuracy and dependability, text and images have been given precedence in depression identification. The analysis of literature demonstrated the lack of evaluation using multimodal data and an optimized hybrid approach which is required to perform in-depth and accurate depression investigation.

Therefore, the present chapter aims to utilize the proposed novel multi-class depression evaluation dataset containing text as well as image data as input to analyze both the Hindi and Hinglish code mixed data by employing a robust hybrid text-image model. The goal is to provide a novel hybrid approach by integrating the best-performing algorithms that produce the final prediction based on earlier individual predictions. Keeping the relevance of both types of data in mind, the work in this study is also conducted on multimodal data by merging BERT and CPSO to develop a unified model that can cope with each type of data. The text and image data are taken in an equal ratio for training and testing purposes.

As described earlier, initially, the CPSO model is proposed that utilizes CNN network where the optimization of the hyperparameters is performed with the help of PSO to classify posts into depressive and non-depressive. Afterwards, BERT and CPSO hybrid is developed for the identification of depressive and non-depressive posts using Hindi and Hinglish language based on both text and image data. The BERT model's text representation is processed using CNN to select the important data characteristics, which are then merged with the image data to provide a more accurate means for diagnosing depression. Thus, an ensemble model, abbreviated as BTCPSO, is developed that can deal with image as well as text data and can perform more accurate predictions. A visual depiction of the proposed novel hybrid approach is presented in Figure 4.9. To get the final class label, the results are fused using a sigmoid activation function. This function outputs a vector detailing the estimated likelihood of each class's output.

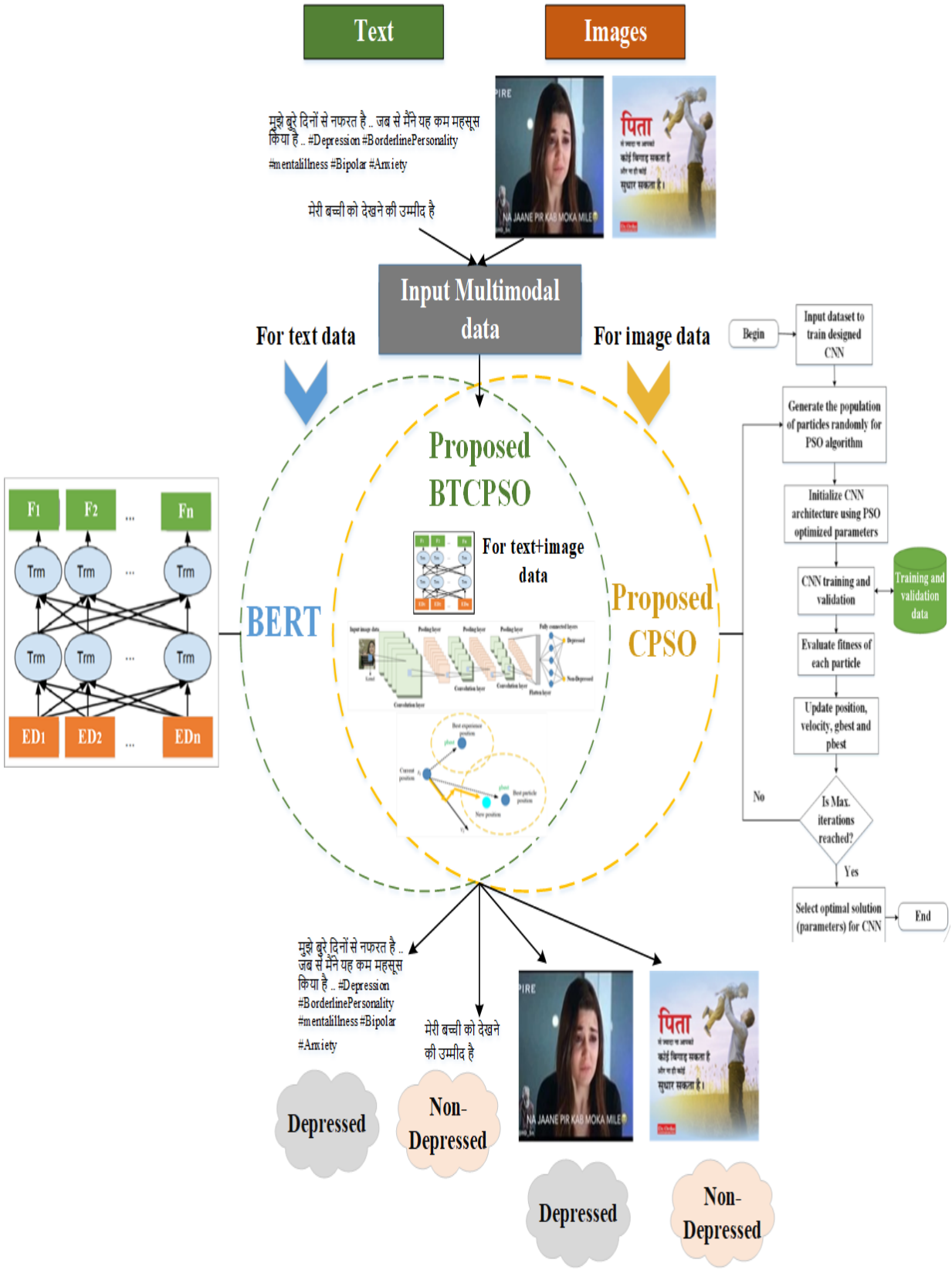


Figure 4.9: Proposed BERT-CPSO (BTCPSO) for multimodal data

4.3 Results and Analysis

This section summarizes the results from various algorithms for text and image categorization employed to diagnose depression. Anaconda Navigator with Python programming platform along with GPU system employing TensorFlow Deep Learning tool and Keras as the library was used for experimentation. The performance evaluation metrics used are accuracy, sensitivity/recall, precision and F1-score which is explained in Chapter 2 Section 2.6.

The entire dataset is divided into the ratio of 70:30 representing the training and testing data. The analysis of results is provided in three experiments: (i) BERT and its variants for text data (ii) Proposed CPSO and other ML as well as DL techniques for image data (iii) Proposed BTCPSO and other combinations for multimodal data. The evaluation of the classification is performed using four metrics as described in the previous section. Before the experimentation the entire dataset has been split in the ratio of 80:10:10 where 80% of the data is used in training the model, 10% validates the model and 10% is used in testing the model. The up arrow (↑) used in the tables denotes that the higher value is better and the down arrow (↓) represents the better results having the lower values.

4.3.1 Results for Text Data Analysis

To perform analysis of Hindi language and Hinglish text data, various Transfer Learning models are employed including BERT, RoBERTa, DistilBERT, and XLNet. The accuracy, recall, precision, and F1-score of the BERT model were found to be 94.21%, 92.14%, 95.32%, and 93.26%, respectively. In addition, the results demonstrate that RoBERTa attained 92.31% accuracy, 90.27% recall, 93.42% precision, and 91.39% F1-score; DistilBERT obtained 90.24% accuracy, 88.31% recall, 91.38% precision, and 89.31% F1-score; and XLNet reached 93.38% accuracy, 91.25% recall, 94.43% precision, and 92.27% F1-score. Therefore, The BERT model outperformed other variants in categorizing pure textual data by a comprehensive evaluation based on assessment metrics. The associated Table 4.6 presents the result values for all models employed for text-based data in which the bold

letter values represent the BERT model, and Figure 4.10 depicts all these graphically.

Table 4.6: Comparative results for text data classification using BERT and its variants

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)
BERT	**94.21%**	**92.14%**	**95.32%**	**93.26%**
RoBERTa	92.31%	90.27%	93.42%	91.39%
DistilBert	90.24%	88.31%	91.38%	89.31%
XLNet	93.38%	91.25%	94.43%	92.27%

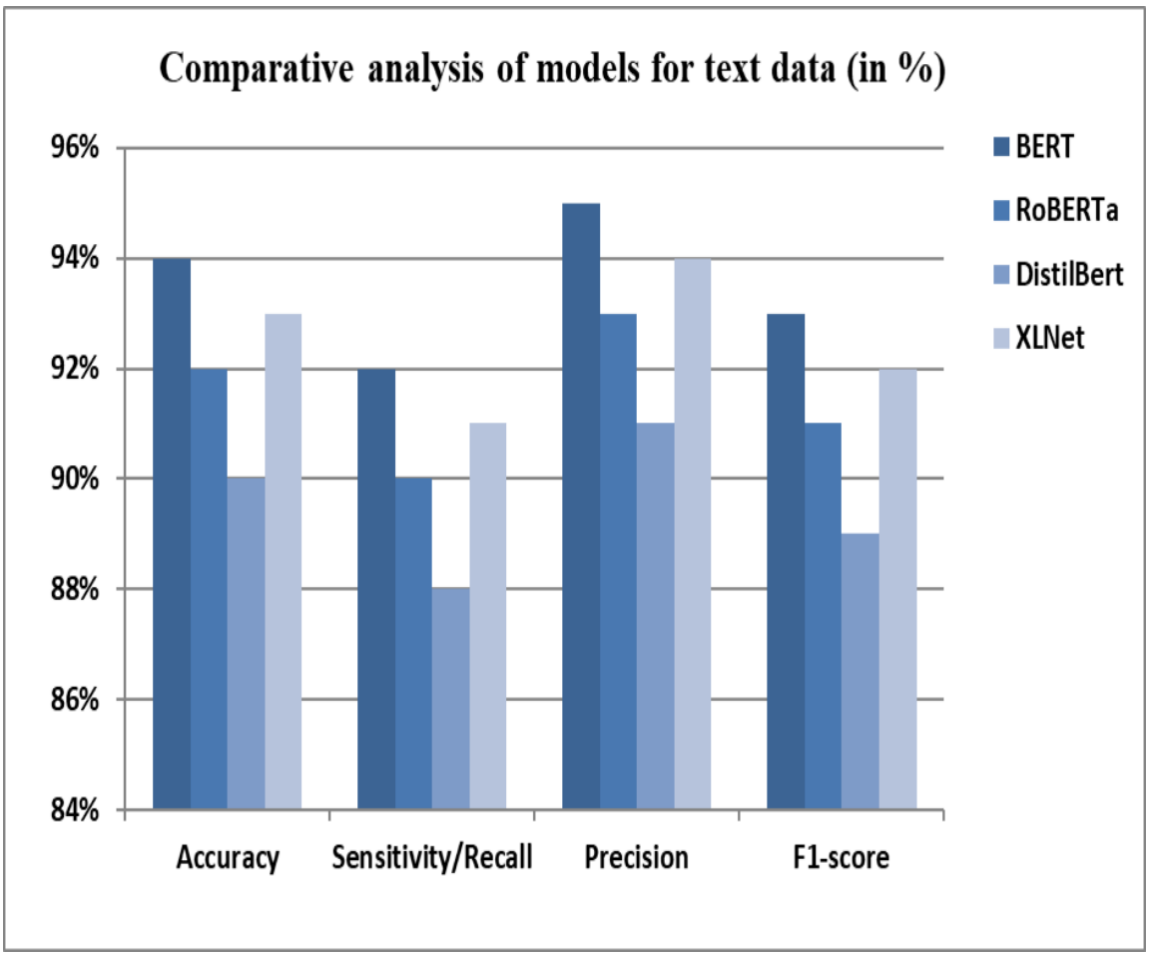


Figure 4.10: Graph showing results for Text data classification using different techniques

4.3.2 Results for Image Data Analysis

To process image data, a model is proposed that integrates CNN and PSO to optimize the hyperparameters of CNN to yield effective and accurate classification. The suggested CPSO model achieved an accuracy of 95.42%, recall of 92.31%, precision of 95.37%, and F1-score of 94.21%, which demonstrate an outstanding performance. By using both ML and DL models in this procedure, the experimentation is expanded. To evaluate their effectiveness, ML models including KNN, SVM, DT, and NB as well as DL methods like ResNet, VGGNet, and MobileNet were used. ML approaches, such as KNN, SVM, DT, and NB, were employed and obtained the accuracy of 85.34%, 88.38%, 80.35%, and 83.39%, the recall of 83.24%, 86.31%, 77.43%, and 80.27%, the precision of 86.29%, 89.17%, 82.34%, and 83.14%, and the F1-score of 84.38%, 87.27%, 79.46%, and 8.18%. Similarly, implementing DL techniques such as ResNet, VGGNet, and MobileNet, the accuracy of 91.23%, 88.13%, and 85.49%, the recall of 89.29%, 86.39%, and 83.25%, the precision of 92.38%, 89.26%, and 86.31%, and the F1-score of 90.41%, 87.31%, and 84.42% were achieved. The effectiveness of the CPSO model surpasses all other approaches by attaining the highest accuracy based on a comparison of results obtained from DL and ML models. The results achieved for the proposed CPSO and other ML and DL approaches are presented in Table 4.7 and Table 4.8 are graphically shown in Figure 4.11 and 4.12.

Table 4.7: Comparative results for image data classification using proposed and ML techniques

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)
CPSO (Proposed)	**95.42%**	**92.31%**	**95.37%**	**94.21%**
KNN	85.34%	83.24%	86.29%	84.38%
SVM	88.38%	86.31%	89.17%	87.27%
DT	80.35%	77.43%	82.34%	79.46%
NB	83.39%	80.27%	83.14%	81.18%

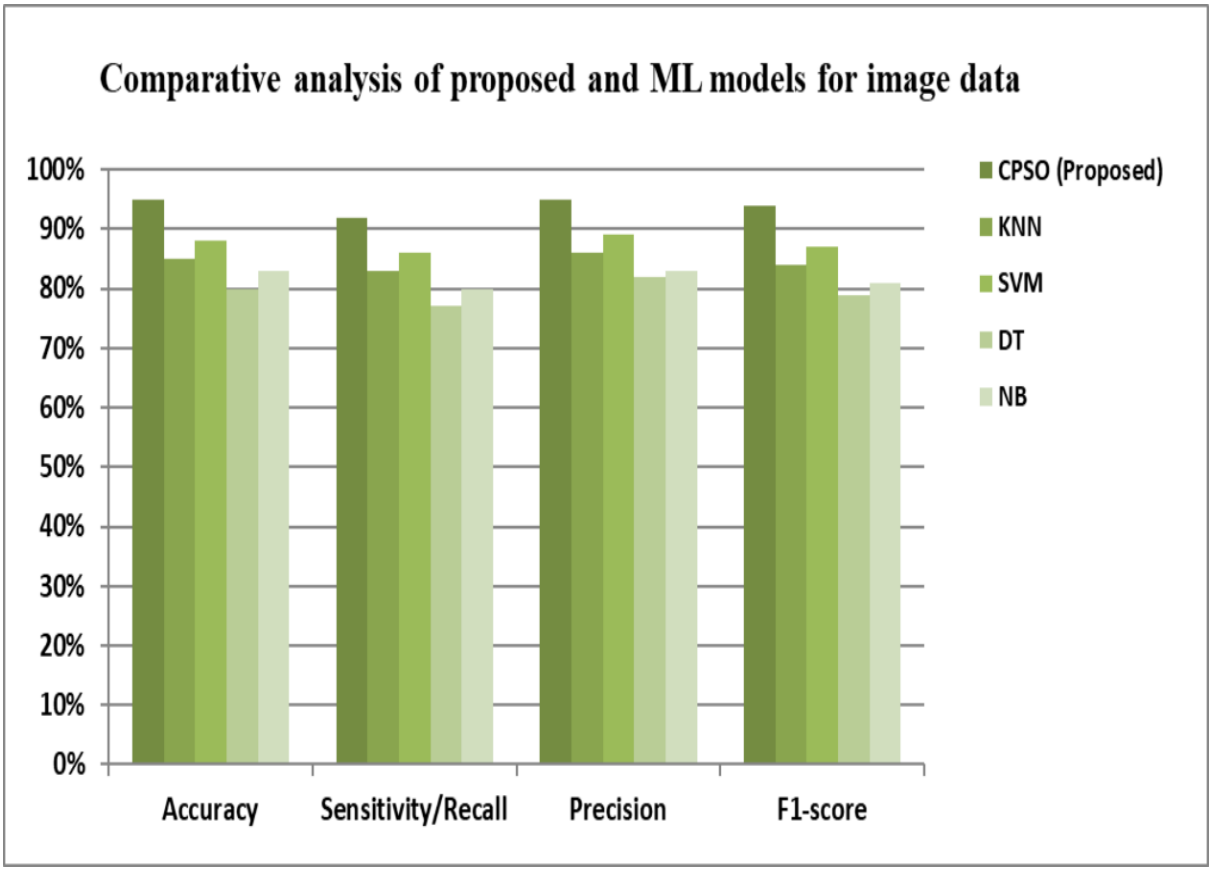


Figure 4.11: Graph showing comparative results for multimodal data classification using CPSO and ML techniques

Table 4.8: Comparative results for image data classification using proposed and DL techniques

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)
CPSO (Proposed)	**95.42%**	**92.31%**	**95.37%**	**94.21%**
ResNet	91.23%	89.29%	92.38%	90.41%
VGGNet	88.13%	86.39%	89.26%	87.31%
MobileNet	85.49%	83.25%	86.31%	84.42%

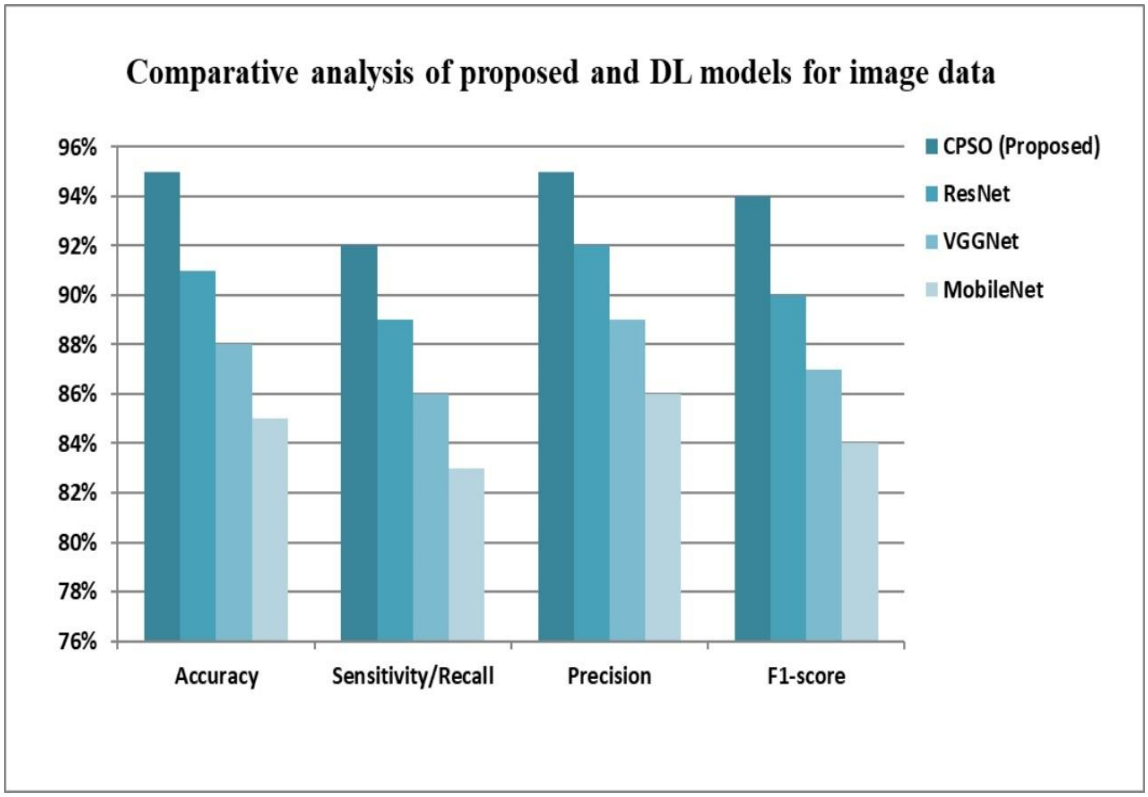


Figure 4.12: Graph showing comparative results for multimodal data classification using CPSO and DL techniques

To further verify the findings, a comparison between the combination of the CNN and PSO (CPSO) and just CNN is performed. The goal was to find out if CNN's hyperparameters optimization with PSO would result in better performance when compared to CNN alone for image classification. The results achieved for this analysis are presented in Table 4.9 and Figure 4.13. depicts the graphical analysis of the obtained results. According to the results from the data, there is a performance improvement when CNN is merged with PSO compared to basic CNN, demonstrating the usefulness of the suggested CPSO technique. Therefore, the hyperparameters using a robust approach can drastically affect the accuracy of the DL model.

Table 4.9: Comparative results for text data classification using proposed and DL techniques

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)
CNN with PSO (CPSO)	95.42%	92.31%	95.37%	94.21%
CNN without PSO	90.27%	88.41%	91.38%	89.26%

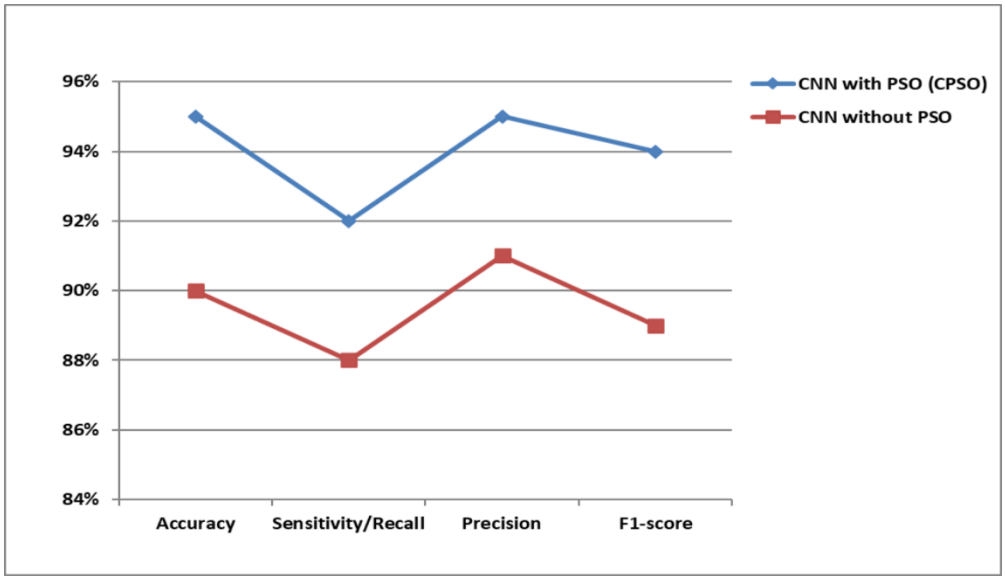


Figure 4.13: Graph showing comparative results for image data classification with and without CNN-PSO combination

4.3.3 Results for Multimodal Data Analysis

The step-by-step process is used to pick the most suitable models for evaluating Hindi language and Hinglish text and image data, i.e. multimodal data, for effective depression detection. To cope with multimodal data, the top-performing models, BERT and CPSO, were chosen. To take benefit of both techniques in this work, a hybrid approach combining BERT and CPSO, i.e. BTCPSO, is presented. Further, to validate the results, the proposed BTCPSO is compared to different BERT combinations such as BERT-KNN, BERT-SVM, BERT-DT, and BERT-NB. The performance assessment metrics for each model are computed, and the findings demonstrated the efficacy of the proposed approach by providing an accuracy of 97.12% in comparison to other ML approaches with the accuracy of 90.18%, 92.33%, 88.37%, and 86.27% for BERT-KNN, BERT-SVM, BERT-DT, and BERT-NB, respectively. Table 4.10 shows the results obtained using the suggested hybrid and different hybrid approaches, revealing the best performance with BTCPSO. Figure 4.14 depicts a graphical analysis of the findings from Table 4.10. Similarly, the results of the proposed BTCPSO are compared with a combination of BERT with DL models

considering BERT-ResNet, BERT-VGGNet, and BERT-MobileNet as presented in Table 4.11, and can be graphically analyzed in Figure 4.15. The data obtained again demonstrated the highest accuracy using the proposed approach for the classification of multimodal data into depressive and non-depressive posts. Also, the error rate for the proposed BTCPSO is comparatively evaluated with other ML and DL approaches. The results yielded revealed the lowest error rate for the proposed approach i.e., 3% in comparison to other learning techniques. Therefore, the proposed hybrid approach can perform accurately within very less error.

Table 4.10 Comparative results for multimodal data classification using proposed and ML techniques

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)	Error(↓)
BTCPSO (Proposed)	**97.12%**	**95.32%**	**98.39%**	**96.46%**	**3%**
BERT-KNN	90.18%	88.32%	91.28%	89.37%	10%
BERT-SVM	92.33%	90.39%	93.26%	91.23%	8%
BERT-DT	88.37%	86.45%	89.32%	87.26%	12%
BERT-NB	86.27%	84.43%	87.22%	85.34%	14%

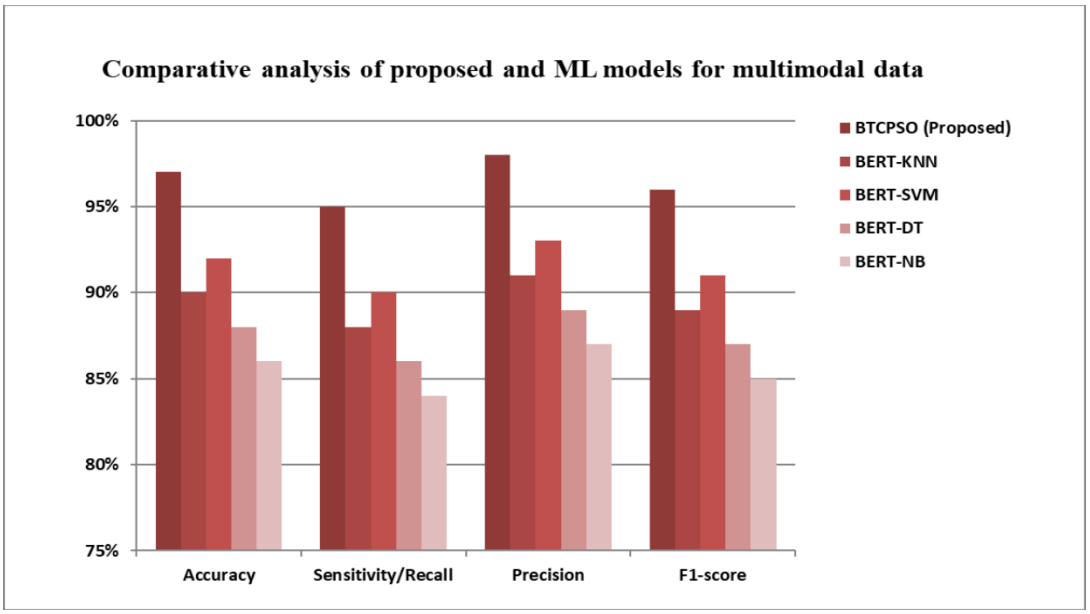


Figure 4.14: Graph showing comparative results for multimodal data classification using proposed and ML techniques

Table 4.11 Comparative results for multimodal data classification using proposed and DL techniques

Model	Accuracy(↑)	Sensitivity/Recall(↑)	Precision(↑)	F1-score(↑)	Error(↓)
BTCPSO (Proposed)	**97.12%**	**95.32%**	**98.39%**	**96.46%**	**3%**
BERT-ResNet	92.34%	90.26%	93.29%	91.31%	8%
BERT-VGGNet	91.35%	89.24%	92.21%	90.32%	9%
BERT-MobileNet	90.14%	88.34%	91.27%	89.37%	10%

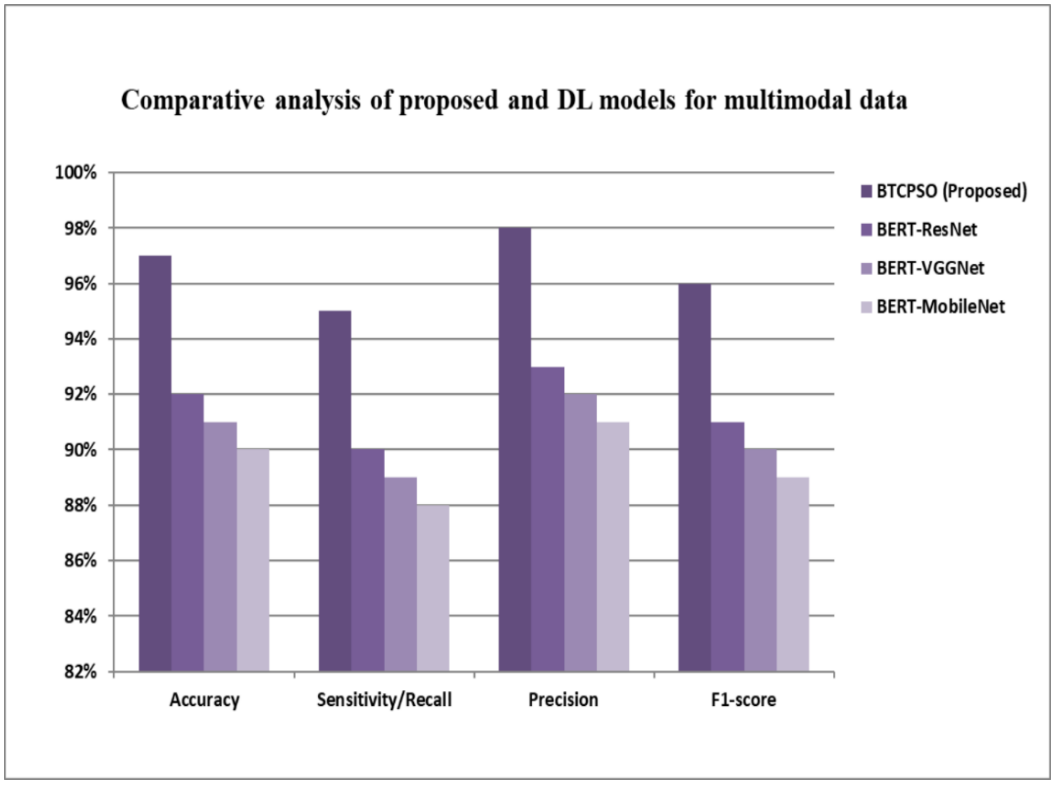


Figure 4.15: Graph showing comparative results for multimodal data classification using proposed and DL techniques

Further, it is clear from Figure 4.14 that BERT-NB produces the lowest results, whereas the suggested BTCPSO technique yields the highest performance values among ML approaches. Similarly, Figure 4.15 indicates that the lowest accuracy is shown by BERT-MobileNet when compared to the proposed BTCPSO. First, the analysis of the classification of text algorithms is done by evaluating each

one to the others. When it comes to classifying texts into depressive and non-depressive categories, BERT showed the best accuracy. Further, the suggested CPSO approach beat the other ML and DL methods in identifying depressive and non-depressive posts from images. Subsequently, a DL-based CNN optimized using the PSO technique is developed and evaluated in comparison to ML and DL algorithms. Finally, the two top-performing models, BERT and CPSO, were hybridized to generate the unified and reliable model (BTCPSO), which performs well with multimodal Hindi language and Hinglish multimodal data. In addition, when compared to the other hybrid models addressed, the proposed hybrid strategy produced the best results with low error.

Atlast, the computational complexity of the proposed BTCPSO is evaluated utilizing Big O Notation. The architecture of DL based proposed technique comprises of multiple hidden layers, h, and requires i number of iterations for training. Thus, the time complexity can be defined as O(ihtno), where t, n and o denote training samples, hidden neurons, and number of neurons in input or output layers respectively. Further, the time complexity of PSO can be computed as O(SzDa), where Sz and Da represent the swarm size, and adjustable parameter's dimension. As the time complexity of Da is O(tno), so, the overall time complexity of PSO and BTCPSO becomes O(Sztno), and O(Szihtno). In addition, the proposed CNN based architecture offers less computational complexity due to its simple architecture in comparison to other employed models. Further, it is analyzed that the time complexity of the DL model varies with increasing and decreasing the number of layers.

In optimization challenges, a hybridized approach that combines CNN with PSO presents several kinds of benefits. The process of training can be improved through the integration of PSO with CNNs, which will help the network identify better sets of weights and biases. In addition, PSO guides the search process, permitting CNNs to arrive at the correct response faster. This can significantly cut the amount of time and computing power required for training Deep Neural Networks. Also, PSO can enable the CNN's hyperparameters to be modified, reducing the possibility that it would overfit the training data. CNNs' chances of finding a global optimum are

boosted by using PSO's exploration-exploitation trade-off to stop them from getting caught in local minima. Thus, the CPSO hybrid technique is adjustable and may be used for a number of applications, including feature selection, detecting objects, and image classification. Furthermore, the BERT model's distinctive capacity to process input in both directions enables the extraction of additional contextual information and increases the model's overall performance. In addition, when BERT, identified for its contextual embedded words, and CNN, known for being able to record local characteristics, are combined, they produce an effective synergy. Incorporating the beneficial features of both BERT and the optimized CNN (CPSO), it creates a model that is efficient and flexible and successfully captures both local and global contexts.

4.3.4 Statistical Analysis

The results of this chapter revealed the best performance by employing the proposed approach. Further, the outcomes are validated using a well-known statistical test Friedman's Rank Test (FRT). To perform the comparative analysis and determine the performance of the proposed and other models, the datasets, [124 & 125], are also experimented using FRT. These datasets comprise of depression-related data in the form of different user's post. The various accuracy rates gained by the suggested and other techniques by employing FRT on the focused datasets is depicted in Table 4.12. The significant variations between the models are confirmed by the FRT. The null hypothesis is created, followed by an assessment on the basis of ranks has been carried out. A null hypothesis claims that there are no significant variations and that all of the methods demonstrate an identical accuracy rate.

Furthermore, the model having the highest possible accuracy rate acquires the highest rank from the FRT and vice-versa. A p-value of 0.0425, which represents a lower value than the significance level of $p<0.05$, was gained for the test results. The null hypothesis was discarded as a consequence of the test findings revealing that there is substantial variation in the performance of models across different datasets. The BERT_VGGNet model obtains the lowest rank, in contrast to the proposed approach

which reflects the best performance achieves the highest rating i.e., rank 4. The FRT results, corresponding p-value, and average rank of the techniques employed are shown in Table 4.13 below.

Table 4.12: Accuracy achieved by models on different datasets

Datasets/Models	**Proposed**	**BERT_ ResNet**	**BERT_ VGGNet**	**BERT_MobileNet**
Proposed	97.12%	92.34%	91.35%	90.14%
[124]	94.35%	88.35%	86.41%	86.36%
[125]	96.25%	90.31%	89.14%	89.37%

Table 4.13: Results of FRT

Rank	**1st**	**2nd**	**3rd**	**4th**	**p-value**
Model	BERT_VGG Net	BERT_Mobile Net	BERT_Res Net	Proposed	0.0425
Average rank	1.3333	1.6667	3	4	

4.3.5 Comparison with state-of-the-art studies

After the proposed and other strategies have been evaluated on the complete dataset, it is vital to compare their performance to earlier studies to verify their success. Therefore, the proposed BTCPSO technique is compared with a few recent, cutting-edge investigations to ascertain how effective it is. Since accuracy and F1-score are frequently employed in all research initiatives for purposes of comparability, they are used as a means of evaluation metric. Table 4.14 offers details about this comparison, which is also shown graphically in Figure 4.16 and Figure 4.17 below. Kumar et al. [64] used Twitter data in the form of English text to analyze depression with the help of different approaches. The evaluation of real-time social platform posts using the ensemble method provided the highest accuracy of 85.09% and F1-score of 79.68%. Similarly, another study by Bhowmik et al. [69] and Khan et

al. [55] performed the analysis of sentiments based on Urdu and Bangla languages. The data was gathered from social media sites in textual form and the accuracy of 78.69% and 82.50% were achieved by the respective studies. Chekima et al. [67] performed the analysis of sentiments based on Malay language after gathering the textual data from social media sites and achieved an accuracy of 79.28%. Carmel et al. [89] proposed a study towards utilizing code-mixed data for depression detection. They evaluated Hindi-English tweets in the form of text and yielded an accuracy of 96.15% and F1-score of 91.4% using the MNB classifier. A very small number of researchers have examined pure Hindi language data, which should be covered according to the analysis of previous research.

Therefore, the evaluation of multimodal pure Hindi language and Hinglish data (text, emoticons, and pictures) was taken into account in the proposed study. Additionally, the lack of a rigorous optimization method for adjusting a DL model's hyperparameters enhanced this research's interest in performing such an analysis.

Table 4.14: Comparison of the proposed approach with previous studies

Reference/Year	**Modality**	**Language**	**Technique**	**Data Source**	**Average Accuracy and F1-score**
Kumar [64]/ 2019	Text	English	NB, RF, GB and Ensemble	Twitter	Highest with Ensemble Accuracy= 85.09% F1-score = 79.68%
Chekima [67]/ 2020	Text	Malay	Proposed model, SVM, NB, ME, and baseline	Facebook and Twitter	Accuracy = 79.28% (with proposed)
Bhowmik [69]/ 2021	Text	Bangla	LR, KNN, RF, SVM	Cricket and Restaurant	Highest with SVM on cricket dataset Accuracy=

					78.69% F1-score = 78.93%
Khan [55]/ 2022	Text	Urdu	Fast Text and BERT (proposed)	UCSA-21	Proposed BERT Accuracy = 82.50% F1-score = 81.49%
Carmel [89]/2022	Text	Hindi-English	MNB algorithm	Twitter	Accuracy = 96.15% F1-score = 91.4%
Proposed Study	Text, images, and emoticons	Pure Hindi and Hinglish	*For text:* BERT, RoBERTa, DistilBERT, and XLNet, *For images:* Proposed CPSO, SVM, DT, KNN, NB, *For (text+images):* Hybrid BTCPSO (proposed) and other combinations	Instagram and other social media posts	Highest with: **BERT:** Accuracy = 94.21%, F1-score = 93.26% **Proposed CPSO:** Accuracy =95.42%, F1-score = 94.21% **Hybrid BTCPSO:** Accuracy = 97.12%, F1-score = 96.46%

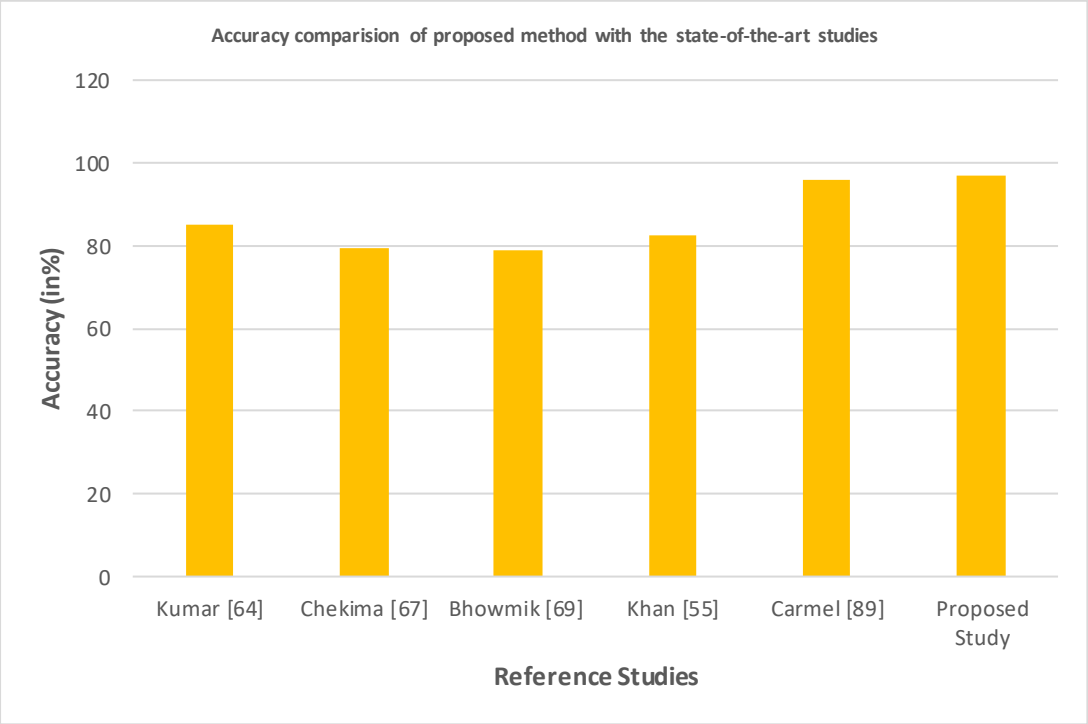


Figure 4.16: Accuracy comparison of the proposed approach with previous studies

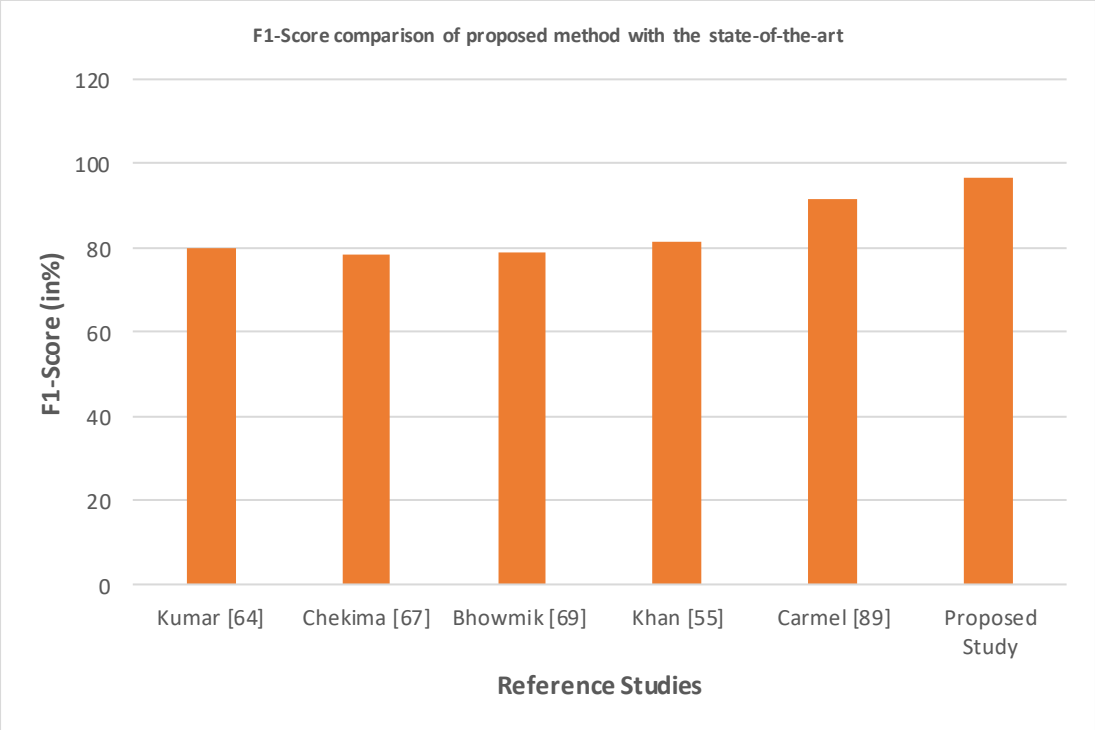


Figure 4.17: F1-score comparison of the proposed approach with previous studies

The optimal selection of CNN hyperparameters using PSO has shown reliable results. The primary objective of this study is to develop a reliable and ideal DL-based technique that can be used for all forms of data to identify posts with and without depression. The results shown in Table 4.14 demonstrate that, with an accuracy of 97.12% and F1-score of 96.46%, the proposed hybrid BTCPSO technique has outperformed the state-of-the-art. Therefore, the proposed framework has a strong potential for handling multimodal Hindi language and Hinglish multimodal data and can predict depression with high performance.

4.4 Chapter Summary

In the era of notable technological developments, online social media platforms have become one of the most prevalent ways for users to convey their views, emotions, ideas, and perspectives via different modalities such as images, text, audio, videos, etc. A thorough evaluation of such information could provide crucial indications about a person's psychological state. Depression seems to be a fast-growing disease, particularly among the younger population globally. Therefore, it is crucial to detect depression early so that people may receive quick counseling and treatment. The objective of this chapter was to build a novel Hindi language and Hinglish-based depression dataset and to develop a unique framework for assessing depression utilizing the created dataset, which has been neglected in prior studies. The objective of this work is threefold: (i) to offer a successful technique for analyzing text data; (ii) to provide an optimized approach to deal with image data using the combination of nature-inspired algorithms and the DL models; and (iii) to propose a high-performance hybrid approach for analyzing multimodal i.e. text as well as image data.

Firstly, relevant data, consisting of pure Hindi and Hinglish, is gathered from various social media platforms, and different approaches are employed to analyze depression. For textual data, BERT is employed due to its unique property of capturing relevant contextual details and a comparison is performed with its variants including RoBERTa, DistilBERT, and XLNet. The analysis of results showed the highest

performance with BERT i.e. accuracy of 94.21% as compared to other variants for text analysis. Furthermore, CNN is improved by using an optimization technique such as PSO to optimize hyperparameters and attain higher results for image data. PSO controls the search process, letting CNNs find the correct answer more quickly. Using PSO's exploration-exploitation trade-off enhances the likelihood of CNNs finding a global optimum. The proposed CPSO approach is also compared to basic CNN to determine the effect of hyperparameter optimization. The collected findings show an increase in accuracy when CNN is optimized using PSO, ranging from 90.27% to 95.42%. In addition, CPSO has been contrasted to other ML and DL algorithms for evaluating efficacy and demonstrated the highest accuracy of all, i.e. 95.42%.

This chapter also provided a unified model to tackle the issue of multimodal data. For such a purpose, the best performing techniques obtained in previous steps for text and image data classification, i.e., BERT and CPSO, are combined to develop a model, namely BTCPSO, that can classify users into depressive and normal based on Hindi language and Hinglish multimodal data. Also, to validate the results, the proposed BTCPSO is compared with other combinations of BERT+ML and BERT+DL techniques. The suggested hybrid BTCPSO approach achieves the highest accuracy of 97.12% when compared to other approaches. As a result, the proposed methodology may be useful for doctors and researchers in the early detection of depression which is aligned with the research objective 3 and bridges the gap of creating a new Hindi/Hinglish (regional language) dataset and creating a model for the same.

CHAPTER 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

This chapter presents the concluding remarks of the thesis by summarizing the key research contributions, outlining the limitations of the work, highlighting its potential social impact, and discussing future directions for further research and development.

5.1 Research Summary

This research focuses on developing the deep learning-based depression detection models using the user generated multimodal social media content. However, there were many constraints regarding the usage of multimodal datasets, availability of the public datasets for depression detection in both English and Hindi languages, usage of robust and unified models and the severity-based depression detection models. To address these challenges, four key research objectives (Ros) were formulated. Each objective has been successfully accomplished through the publication of research articles, as summarized in Table 7.1.

In order to accomplish RO1, a detailed and systematic literature survey of the depression detection models is presented. The objective of this review is fivefold; Initially, the modalities being widely utilized in past works for depression detection are comprehensively analyzed. Secondly, the datasets which are available and used by researchers are studied. Thirdly, the learning techniques which have been utilized in recent years are discovered. Fourthly, the languages which have been practiced in current years are explored. Lastly, the research gaps are identified that need to be covered in future to enable early depression detection. Each of the aspect is scrutinized in detail to determine the extent of work already done. This systematic survey lays a compact foundation for developing deep learning frameworks using multimodal social

media content for depression detection tailored to the research gaps found.

Table 5.1: Research objectives and their corresponding publications

Research Objectives	List of Publication
RO1. To perform the systematic literature review of the multimodal deep learning models for depression detection from social media posts.	1. Saraswat, P., & Beniwal, R. A Systematic Survey of Depression Detection: Modalities, Datasets, and Learning Techniques, 2nd International Conference on Computing and Machine Learning (CML 2025), Sikkim Manipal University, India. Springer. **[Accepted and Presented March 2025]**
RO2. To develop a multimodal deep learning-based framework for detecting depression by affective analysis of Social Media posts from different platforms for Multimedia Dataset.	1. Beniwal, R., & Saraswat, P. A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data, The Computer Journal, Oxford University Press. **[Paper Published – January 2024]** 2. Saraswat, P., & Beniwal, R. BERT Based RNN for Effective Detection of Depression with Severity Levels from Text Data, IEEE Symposium on Wireless Technology and Applications (ISWTA-2024), Malaysia. **[Paper Published – July 2024]**
RO3. To develop a multimodal deep learning-based framework for detecting depression by affective analysis of Social Media posts for the Hindi language content.	1. Beniwal, R., & Saraswat, P. A Hybrid BERT-CPSO Model for Multi-class Depression Detection using Pure Hindi and Hinglish Multimodal Data on Social Media, Computers and Electrical Engineering, Elsevier. **[Paper Published - October 2024]**
RO4. To do a comparative result analysis of the developed models with other existing models/techniques.	1. Beniwal, R., & Saraswat, P. A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data, The Computer Journal, Oxford University Press. **[Paper Published – January 2024]** 2. Saraswat, P., & Beniwal, R. BERT Based RNN for Effective Detection of Depression with Severity Levels from Text Data, IEEE Symposium on Wireless Technology and Applications (ISWTA-2024), Malaysia. **[Paper Published – July 2024]**

	3. Beniwal, R., & Saraswat, P. A Hybrid BERT-CPSO Model for Multi-class Depression Detection using Pure Hindi and Hinglish Multimodal Data on Social Media, Computers and Electrical Engineering, Elsevier. **[Paper Published - October 2024]**

RO2 was attained by developing two models: Primarily for depression detection using English social media content a dataset is created with text, emoticons, and image data, and then preprocessing is performed using diverse techniques. The proposed model (Hybrid BERT-CNN) in the research consists of three parts: first is BERT model, which is trained on only text data also emoticons are converted into a textual form for easy processing; second is CNN which is trained only on image data and the third is the combination of best-performing models i.e., hybrid of BERT and CNN (BERT-CNN) to work on both the text and images with enhanced accuracy. Finally, the hybrid approach is compared with other combinations and previous studies on the basis of performance measures like; accuracy, sensitivity/recall, precision and F1-score for the categorization of user's post into depressive and non-depressive based on multimodal data.

Furthermore, a deep learning-based model was proposed for severity-based depression detection using English social media content. The model performed a multi-class classification utilizing a publicly available text dataset to categorize users into non-depressed, moderately, and severely depressed. A hybrid technique that robustly combines BERT with two versions of RNN i.e., LSTM and GRU. The BERT model is implemented to extract useful word embeddings which are then fed to RNN comprising LSTM, dropout, and GRU layers. The findings revealed that the employed approach outperformed all other employed models and techniques in the past studies.

In order to achieve RO3 we developed a Hindi dataset and suggested reliable methods for depression detection based on multimodal data, i.e., text and images, using the Hindi and Hinglish languages. Three things were accomplished:

first, it evaluated text data using an effective BERT approach and compare it with other transfer learning variants; second, it analyzed image data by presenting a CNN optimized with a nature-inspired algorithm, namely PSO, or CPSO; and third, it classified the multimodal data into depressive and non-depressive posts by presenting a hybrid of the best-performing models on text and images, namely BERT-CPSO. The results produced with the BERT model were compared with other transfer learning models like, RoBERTa, DistilBERT, and XLNet. Further, CPSO outperformed other ML and DL algorithms for image data. Additionally, comparison of the proposed CPSO with a basic CNN revealed that integrating the PSO technique with CNN increased the model's accuracy in detecting depressed posts. Finally proposed hybrid BERT-CPSO outperformed other BERT combinations with ML and DL algorithms for multimodal data on the basis of the performance measure used like; accuracy, recall, precision, and F1-scores.

In order to attain RO4, all three proposed models were compared with the other TL, DL, ML and combinatory models. Also, all the proposed models were also compared with the existing state-of-the-art techniques and previous studies. The results demonstrated a significant improvement in the models in regard to the performance measures utilized. The results consistently demonstrated that the proposed models out-performed the existing state-of-the-art studies and the traditional approaches confirming the efficiency of the models.

5.2 Limitations of the Work

Despite the fact that the proposed models showcased significant performance by outperforming the existing state-of-the-art studies and the traditional methods certain limitations remains:

a. **Dataset Constraint:** The diversity of the data still remains limited because of the possibility of affecting the data through cultural contexts and other regional

languages even after creating English and Hindi datasets.

b. **Dataset for Severity Based Detection:** The severity-based model was designed with the help of publicly available dataset, so multimodality in the data is still missing.

c. **Interpretability of Models:** Though deep learning models were accurate but there is a lack of transparency, becasure the interpretability techniques were not explored in this research like UME and SHAP.

d. **Applicability in Real-time:** This research has not been validated in real time scenarios of social media networks.

5.3 Social Impact

The proposed research has the potential to deliver significant social implications in terms of mental health awareness, early detection and intervention as follows:

a) **Early Diagnosis and Support:** The proposed models can help in early diagnosis of depression and leading to intervention can possibly save lives.

b) **Breach of Language Barriers:** By creating the Hindi and Hinglish dataset alogwith English language dataset the research promotes inclusivity and diversity for multilingual societies, as we live in a diverse civilization.

c) **Stigma Reduction with the help of Advanced Technology:** As there is a stigma attached with the mental health, people avoid asking for help or even discussing with other. But with the help of the advanced technologies preliminary screening can be done though these models and awareness can be spread for people to seek for help.

d) **Public Health Support System:** This research can be beneficial for the mental health support professionals for keeping a track on person's mental health and providing necessary support when required.

5.4 Future Scope

The future scope of the research presented in this thesis is promising and widespread. While both English and Hindi datasets have been created but it can be expanded for underrepresented languages and regional dialects which will ensure the generalizability in the models. Furthermore, integration of explainable AI (XAI) will make the models more realistic and reliable which can be help for mental health support professionals. Moreover, along with text, image and emoticons other modalities can be included altogether like behaviour signals, audio and video.

Additionally, future research can focus on developing real-time models that will learn continuously and evolve with the users by learning their linguistic trends, behaviour and platforms they use.

BIBLIOGRAPHY

[1] Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Hosseini Ghomi, R. (2019). Depression screening from voice samples of patients affected by parkinson's disease. Digital biomarkers, 3(2), 72-82.

[2] PY, K., Dube, R., Barbade, S., Kulkarni, G., Konda, N., & Konkati, M. (2021, May). Depression Detection using Machine Learning. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021).

[3] Julian, L. J. (2011). Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis Care & Research, 63*(S11), S454–S466.

[4] Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales (DASS).* Psychology Foundation of Australia.

[5] Arroll, B., Smith, F. G., Kerse, N., Fishman, T., & Gunn, J. (2005). Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. Bmj, 331(7521), 884.

[6] Ismail, L., Shahin, N., Materwala, H., Hennebelle, A., & Frermann, L. (2023). ML-SocMedEmot: Machine Learning Event-based Social Media Emotion Detection Proactive Framework Addressing Mental Health: A Novel Twitter Dataset and Case Study of COVID-19.

[7] Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. Multimedia Tools and Applications, 81(17), 23649-23685.

[8] WHO highlights urgent need to transform mental health and mental health care. (2022). Retrieved 5 May 2022, from https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care

[9] Mental health. (2022). Retrieved 8 May 2022, from https://www.who.int/india/health-topics/mental-health

[10] Depression. (2022). Retrieved 10 May 2022, from https://www.who.int/news-room/fact-sheets/detail/depression

[11] Benazzi, F. (2006). Various forms of depression. *Dialogues in clinical neuroscience*, *8*(2), 151-161.

[12] Ainsworth, P. (2009). *Understanding depression*. Univ. Press of Mississippi.

[13] Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *Archives of General Psychiatry*. 2000;57(3):241-247. doi:10.1001/archpsyc.57.3.241.

[14] Pariante CM, Lightman SL. The HPA axis in major depression: classical theories and new developments. *Nature Reviews Neuroscience*. 2008;9(10): 687–701. doi:10.1038/nrn2345.

[15] Miller AH, Raison CL. The role of inflammation in depression: from evolutionary imperative to modern treatment target. *Nature Reviews Immunology*. 2016;16(1):22–34. doi:10.1038/nri.2015.5.

[16] Beck AT. *Depression: Clinical, Experimental, and Theoretical Aspects*. New York: Harper & Row; 1967.

[17] Kotov R, Gamez W, Schmidt F, Watson D. Linking "Big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological Bulletin*. 2010;136(5):768–821. doi:10.1037/a0020327.

[18] Heim C, Nemeroff CB. The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies. *Biological Psychiatry*. 2001;49(12):1023–1039. doi:10.1016/S0006-3223(01)01157-X.

[19] Seligman ME. Learned helplessness. *American Psychologist*. 1975;30(4): 406–420. doi:10.1037/h0077357.

[20] Kendler KS, Karkowski LM, Prescott CA. Stressful life events and previous episodes in the etiology of major depression in women: an evaluation of the "kindling" hypothesis. *American Journal of Psychiatry*. 1999;156(6): 837–841. doi:10.1176/ajp.156.6.837.

[21] Cacioppo JT, Hawkley LC. Perceived social isolation and cognition. *Trends in Cognitive Sciences*. 2009;13(10):447–454. doi:10.1016/j.tics.2009.06.005.

[22] Lorant V, Deliège D, Eaton W, Robert A, Philippot P, Ansseau M. Socioeconomic inequalities in depression: a meta-analysis. *British Journal of Psychiatry*. 2003;187(6): 457–463. doi:10.1192/bjp.187.6.457.

[23] Yang LH, Kleinman A, Link BG, Phelan JC, Lee S, Good B. Culture and stigma: adding moral experience to stigma theory. *Transcultural Psychiatry*. 2007;44(4):475–495. doi:10.1177/1363461507084101.

[24] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 5th ed.* Arlington, VA: American Psychiatric Publishing; 2013.

[25] World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. Geneva: WHO; 2017.

[26] First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015). *Structured Clinical Interview for DSM-5 Disorders, Clinician Version (SCID-5-CV).* American Psychiatric Association Publishing.

[27] Jørgensen, H., & Bech, P. (2011). Rating scales in depression: limitations and pitfalls. *Dialogues in Clinical Neuroscience, 13*(1), 93–100.

[28] Feldman, L., Zhang, Y., & Singh, R. (2023). *Artificial Intelligence-Enabled Wearables for Depression and Anxiety Detection: A Meta-Analysis*. Frontiers in Digital Health, **5**, 1012345.

[29] Every hour, one student commits suicide in India. (2022). Retrieved 10 May 2022, from https://www.hindustantimes.com/health-and-fitness/every-hour-one-student-commits-suicide-in-india/story-7UFFhSs6h1HNgrNO60FZ2O.html

[30] Gurgaon man 'commits suicide' live on Facebook: Police. (2022). Retrieved 10 May 2022, from https://indianexpress.com/article/india/gurgaon-man-commits-suicide-live-on-facebook-police-5287059/

[31] 'I am unable to live anymore', man goes live on social media before ending life; kin booked. (2022). Retrieved 10 May 2022, from https://www.timesnownews.com/ahmedabad/article/gujarat-morbi-man-suicide-live-social-media-harassed-by-in-laws/795305

[32] Facebook: 24-year-old Agra youth live-streams suicide on Facebook. (2022). Retrieved 10 May 2022, from https://economictimes.indiatimes.com/news/politics-and-nation/24-year-old-agra-youth-live-streams-suicide-on-facebook/articleshow/64957797.cms?from=mdr

[33] Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying mental health signals in Twitter. In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality (pp. 51-60).

[34] Lin, H., Jia, J., Nie, L., Shen, G., & Chua, T. S. (2016, July). What Does Social Media Say about Your Stress?. In IJCAI (pp. 3775-3781).

[35] Akbari, M., Hu, X., Liqiang, N., & Chua, T. S. (2016, February). From tweets to wellness: Wellness event detection from twitter streams. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[36] Mahnken, K. (2022). Survey: More Young People Are Depressed During the Pandemic. But They May Be Using Social Media to Cope. Retrieved 11 May 2022, from https://www.the74million.org/survey-more-young-people-are-depressed-during-the-pandemic-but-they-may-be-using-social-media-to-cope/

[37] Hawn, C. (2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. Health affairs, 28(2), 361-368.

[38] Neuhauser, L., & Kreps, G. L. (2003). Rethinking communication in the e-health era. Journal of health psychology, 8(1), 7-23.

[39] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, March). Identifying health-related topics on twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 18-25). Springer, Berlin, Heidelberg.

[40] Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 38(3), 182-188.

[41] Aleem, S., Huda, N. U., Amin, R., Khalid, S., Alshamrani, S. S., & Alshehri, A. (2022). Machine learning algorithms for depression: diagnosis, insights, and research directions. Electronics, 11(7), 1111.

[42] Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications, 34(1), 721-744.

[43] Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Transactions on Biomedical Engineering, 63(3), 664-675.

[44] Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. Journal of machine learning research, 3(Aug), 115-143.

[45] Malhotra, A., & Jindal, R. (2020). Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. EAI Endorsed Transactions on Pervasive Health and Technology, 6(21), e1.

[46] Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine, 135, 104499.

[47] Mo, M., Khan, A., & Gul, N. A Hybrid Method for Stress and Depression Detection from Social Media Text Focused on Local Languages.

[48] Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications, 34(1), 721-744.

[49] William, D., & Suhartono, D. (2021). Text-based depression detection on social media posts: A systematic literature review. Procedia Computer Science, 179, 582-589.

[50] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science, 167, 1258-1267.

[51] Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25(1), 281-304.

[52] Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications, 78(17), 24103-24119.

[53] Solakidis, G. S., Vavliakis, K. N., & Mitkas, P. A. (2014, August). Multilingual sentiment analysis using emoticons and keywords. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (Vol. 2, pp. 102-109). IEEE.

[54] Vandana., Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. Measurement: Sensors, 25, 100587.

[55] Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of urdu text using multilingual BERT. Scientific Reports, 12(1), 5436.

[56] Goodyear, V. A., Wood, G., Skinner, B., & Thompson, J. L. (2021). The effect of social media interventions on physical activity and dietary behaviours in young people and adults: a systematic review. International Journal of Behavioral Nutrition and Physical Activity, 18(1), 72.

[57] Biradar, S., Saumya, S., & Chauhan, A. (2021, December). Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2470-2475). IEEE.

[58] Du, M., Liu, S., Wang, T., Zhang, W., Ke, Y., Chen, L., & Ming, D. (2023). Depression recognition using a proposed speech chain model fusing speech production and perception features. Journal of Affective Disorders, 323, 299-308.

[59] Chlasta, K., Wołk, K., & Krejtz, I. (2019). Automated speech-based screening of depression using deep convolutional neural networks. Procedia Computer Science, 164, 618-628.

[60] Salekin, A., Eberle, J. W., Glenn, J. J., Teachman, B. A., & Stankovic, J. A. (2018). A weakly supervised learning framework for detecting social anxiety and depression. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2(2), 1-26.

[61] Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Uddin, M. (2022). Deep learning for depression detection from textual data. Electronics, 11(5), 676.

[62] Gupta, S., Goel, L., Singh, A., Prasad, A., & Ullah, M. A. (2022). Psychological analysis for depression detection from social networking sites. Computational Intelligence and Neuroscience, 2022.

[63] Arora, P., & Arora, P. (2019, March). Mining twitter data for depression detection. In 2019 international conference on signal processing and communication (ICSC) (pp. 186-189). IEEE.

[64] Kumar, A., Sharma, A., & Arora, A. (2019). Anxious depression prediction in real-time social data. arXiv preprint arXiv:1903.10222.

[65] Singh, D. (2021). Detection of emotions in hindi-english code mixed text data. arXiv preprint arXiv:2105.09226.

[66] Chopra, A., Sharma, D. K., Jha, A., & Ghosh, U. (2023). A Framework for Online Hate Speech Detection on Code-mixed Hindi-English Text and Hindi Text in Devanagari. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(5), 1-21.

[67] Chekima, K., & Alfred, R. (2018). Sentiment analysis of Malay social media text. In Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017 (pp. 205-219). Springer Singapore.

[68] Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., & Gelbukh, A. (2020). A multiclass depression detection in social media based on sentiment analysis. In 17th International Conference on Information Technology–New Generations (ITNG 2020) (pp. 659-662). Springer International Publishing.

[69] Bhowmik, N. R., Arifuzzaman, M., Mondal, M. R. H., & Islam, M. S. (2021). Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. *Natural Language Processing Research*, *1*(3-4), 34-45.

[70] de Melo, W. C., Granger, E., & Hadid, A. (2019, May). Combining global and local convolutional 3d networks for detecting depression from facial expressions. In 2019 14th ieee international conference on automatic face & gesture recognition (fg 2019) (pp. 1-8). IEEE.

[71] Al Jazaery, M., & Guo, G. (2018). Video-based depression level analysis by encoding deep spatiotemporal features. IEEE Transactions on Affective Computing, 12(1), 262-268.

[72] Uddin, M. A., Joolee, J. B., & Lee, Y. K. (2020). Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. IEEE Transactions on Affective Computing, 13(2), 864-870.

[73] Chao, L., Tao, J., Yang, M., & Li, Y. (2015, September). Multi task sequence learning for depression scale prediction from video. In 2015 International conference on affective computing and intelligent interaction (ACII) (pp. 526-531). IEEE

[74] Kamalesh, M. D., & B, B. (2022). Personality prediction model for social media using machine learning technique. Computers and Electrical Engineering, 100, 107852.

[75] Wu, M. Y., Shen, C. Y., Wang, E. T., & Chen, A. L. (2020). A deep architecture for depression detection using posting, behavior, and living environment data. Journal of Intelligent Information Systems, 54, 225-244.

[76] Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92-104.

[77] Han, Y., Liu, M., & Jing, W. (2020). Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. IEEE Access, 8, 21314-21325.

[78] Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9, 1-23.

[79] Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., & Kaewpitakkun, Y. (2018, July). Facebook social media for depression detection in the Thai community. In 2018 15th international joint conference on computer science and software engineering (jcsse) (pp. 1-6). IEEE

[80] Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., ... & Feng, L. (2014, November). User-level psychological stress detection from social media using deep neural network. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 507-516).

[81] Deshpande, M., & Rao, V. (2017, December). Depression detection using emotion artificial intelligence. In 2017 international conference on intelligent sustainable systems (iciss) (pp. 858-862). IEEE.

[82] Das, A. K., Al Asif, A., Paul, A., & Hossain, M. N. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems, 30(1), 578-591.

[83] Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J. (2013, October). Diagnosis of depression by behavioural signals: a multimodal approach. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (pp. 11-20).

[84] Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F., & Turabzadeh, S. (2014, November). Automatic depression scale prediction using facial expression dynamics and regression. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 73-80).

[85] A. Anshul, G. S. Pranav, M. Z. U. Rehman and N. Kumar, A Multimodal Framework for Depression Detection During COVID-19 via Harvesting Social Media, in *IEEE Transactions on Computational Social Systems, vol. 11, no. 2, pp. 2872-2888, April 2024, doi: 10.1109/TCSS.2023.3309229.*

[86] Sadeghi, Misha & Richer, Robert & Egger, Bernhard & Schindler-Gmelch, Lena & Rupp, Lydia & Rahimi, Farnaz & Berking, Matthias & Eskofier, Bjoern. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research. 3. 10.1038/s44184-024-00112-8.

[87] Han, Z., Shang, Y., Shao, Z., Liu, J., Guo, G., Liu, T., ... & Hu, Q. (2023). Spatial–Temporal Feature Network for Speech-Based Depression Recognition. IEEE Transactions on Cognitive and Developmental Systems, 16(1), 308-318.

[88] Nadeem, A., Naveed, M., Islam Satti, M., Afzal, H., Ahmad, T., & Kim, K. I. (2022). Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data. Sensors, 22(24), 9775.

[89] MJ, C. M. B., Arif, M., V, D. K., & K, A. K. (2022). Linguistic Analysis of Hindi-English Mixed Tweets for Depression Detection. Journal of Mathematics, 2022, 1-7.

[90] Poświata, R., & Perełkiewicz, M. (2022, May). OPI@ LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 276-282).

[91] Dai, Z., Zhou, H., Ba, Q., Zhou, Y., Wang, L., & Li, G. (2021). Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. Journal of affective disorders, 295, 1040-1048.

[92] Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy, 22(6), 688.

[93] Wang, Y., Wang, Z., Li, C., Zhang, Y., & Wang, H. (2020). A multitask deep learning approach for user depression detection on sina weibo. arXiv preprint arXiv:2008.11708.

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

[95] Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. Ieee Access, 8, 138162-138169.

[96] Uddin, 0A. H., Bapery, D., & Arif, A. S. M. (2019, July). Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

[97] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 6, 1-12.

[98] Chen, B., Huang, Q., Chen, Y., Cheng, L., & Chen, R. (2018, June). Deep neural networks for multi-class sentiment classification. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th

International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 854-859). IEEE.

[99] Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. EPJ Data Science, 6(1), 15.

[100] Almaev, T. R., & Valstar, M. F. (2013, September). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 356-361). IEEE.

[101] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. In Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DM Apps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers 17 (pp. 201-213). Springer Berlin Heidelberg.

[102] Wu, M. Y., Shen, C. Y., Wang, E. T., & Chen, A. L. (2020). A deep architecture for depression detection using posting, behavior, and living environment data. Journal of Intelligent Information Systems, 54, 225-244.

[103] Abaker, A. A., & Saeed, F. A. (2021). A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. Informatica, 45(1).

[104] AlSagri, H. S., & Ykhlef, M. (2020). Machine learning-based approach for depression detection in twitter using content and activity features. IEICE Transactions on Information and Systems, 103(8), 1825-1832.

[105] Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social media and suicide: a public health perspective. American journal of public health, 102(S2), S195-S200.

[106] Learning, D. (2020). Deep learning. High-dimensional fuzzy clustering.

[107] https://github.com/PaviSaraswat/Socialmedia_DepressionandNondepressive_Imageandtext_data_English_language

[108] Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8(1), 1-20.

[109] Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment analysis-strategy for text pre-processing. Int. J. Comput. Appl, 180(34), 15-18.

[110] Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., & Manicardi, S. (2016, September). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb.

[111] Jana, R., Chowdhury, A. R., & Islam, M. (2014). Optical character recognition from text image. International Journal of Computer Applications Technology and Research, 3(4), 240-244.

[112] Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. International Journal of Advanced Computer Science and Applications, 10(7).

[113] Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences, 9(19), 4062.

[114] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120.

[115] Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021, August). A robustly optimized BERT pre-training approach with post-training. In Proceedings of the 20th chinese national conference on computational linguistics (pp. 1218-1227).

[116] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[117] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[118] He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics, 83, 103-111.

[119] Kim, A. Y., Jang, E. H., Lee, S. H., Choi, K. Y., Park, J. G., & Shin, H. C. (2023). Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach. Journal of Medical Internet Research, 25, e34474.

[120] Samet, H. (2007). K-nearest neighbor finding using MaxNearestDist. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), 243-252.

[121] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

[122] Rokach, L., & Maimon, O. (2005). Decision trees. Data mining and knowledge discovery handbook, 165-192.

[123] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[124] P.-J. Yang. (2020). Detect Depression in Twitter Posts. [Online]. Available: https://github.com/peijoy/DetectDepressionInTwitterPosts

[125] P. Girish. (2021). Depression Detection Using MLAlgorithm. [Online]. Available: https://github.com/patilgirish815/Depression_Detection_Using_ML

[126] Ramalingam, D., Sharma, V., & Zar, P. (2019). Study of depression analysis using machine learning techniques. Int. J. Innov. Technol. Explor. Eng, 8(7C2), 187-191.

[127] Li, X., Zhang, X., Zhu, J., Mao, W., Sun, S., Wang, Z., ... & Hu, B. (2019). Depression recognition using machine learning methods with different feature generation strategies. Artificial intelligence in medicine, 99, 101696.

[128] Teens Spend Average of 4.8 Hours on Social Media Per Day. (2023). Retrieved 5 April 2025. Available (online): https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx

[129] Janatdoust, M., Ehsani-Besheli, F., & Zeinali, H. (2022, May). KADO@ LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 265-269).

[130] Thati, R. P., Dhadwal, A. S., Kumar, P., & P, S. (2023). A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms. Multimedia Tools and Applications, 82(4), 4787-4820.

[131] Lau, C., Zhu, X., & Chan, W. Y. (2023). Automatic depression severity assessment with deep learning using parameter-efficient tuning. Frontiers in Psychiatry, 14, 1160291.

[132] Muñoz, S., & Iglesias, C. Á. (2023). Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representati ons. Applied Sciences, 13(21), 11695.

[133] Stepanov, E. A., Lathuiliere, S., Chowdhury, S. A., Ghosh, A., Vieriu, R. L., Sebe, N., & Riccardi, G. (2018, September). Depression severity estimation from multiple modalities. In 2018 ieee 20th international conference on e-health networking, applications and services (healthcom) (pp. 1-6). IEEE.

[134] Ilias, L., Mouzakitis, S., & Askounis, D. (2023). Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media. IEEE Transactions on Computational Social Systems.

[135] Kayalvizhi, S., Durairaj, T., & Chakravarthi, B. R. (2022, May). Findings of the shared task on detecting signs of depression from social media. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 331-338).

[136] Kayalvizhi, S., Durairaj, T., & Chakravarthi, B. R. (2022, May). Findings of the shared task on detecting signs of depression from social media. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 331-338).

[137] Bhat, P., Anuse, A., Kute, R., Bhadade, R. S., & Purnaye, P. (2022). Mental health analyzer for depression detection based on textual analysis. Journal of Advances in Information Technology Vol, 13(1).

[138] Ullah, N., Khan, J. A., Khan, M. S., Khan, W., Hassan, I., Obayya, M., ... & Salama, A. S. (2022). An effective approach to detect and identify brain tumors using transfer learning. Applied Sciences, 12(11), 5645.

[139] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[140] Shreyashree, S., Sunagar, P., Rajarajeswari, S., & Kanavalli, A. (2022). BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings. International Journal of Advanced Computer Science and Applications, 13(9).

[141] Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In Twelfth Annual Conference of the International Speech Communication Association.

[142] Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, *2*(2), 1-31.

[143] Lee, J. (2020). Mental health effects of school closures during COVID-19. *The Lancet Child & Adolescent Health*, *4*(6), 421.

[144] Kolajo, T., Daramola, O., Adebiyi, A., & Seth, A. (2020). A framework for pre-processing of social media feeds based on integrated local knowledge base. *Information processing & management*, *57*(6), 102348.

[145] Singh, R., Choudhary, N., & Shrivastava, M. (2018, March). Automatic normalization of word variations in code-mixed social media text. In *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 371-381). Cham: Springer Nature Switzerland.

[146] Jamil, M. L., Pais, S., Cordeiro, J., & Dias, G. (2022). Detection of extreme sentiments on social networks with BERT. *Social Network Analysis and Mining*, *12*(1), 55.

[147] Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *Ieee Access*, *7*, 154290-154299.

[148] Yang, X. S., & Karamanoglu, M. (2013). Swarm intelligence and bio-inspired computation: an overview. *Swarm intelligence and bio-inspired computation*, 3-23.

[149] Das, S., Abraham, A., & Konar, A. (2008). Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives. *Advances of computational intelligence in industrial systems*, 1-38.

[150] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. *Encyclopedia of machine learning*, *15*(1), 713-714.

[151] Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

[152] Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019, July). Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 110-117).

[153] Wang, W., Li, Y., Zou, T., Wang, X., You, J., & Luo, Y. (2020). A novel image classification approach via dense-MobileNet models. *Mobile Information Systems*, *2020*.

Appendix A

List of Publications with Proof

Journal Publication

[1] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CNN approach for depression detection on social media using multimodal data. *The Computer Journal*, *67*(7), 2453-2472. https://doi.org/10.1093/comjnl/bxae018, **[Paper Published – January 2024] [SCIE Indexed]**

[2] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CPSO model for multi-class depression detection using pure hindi and hinglish multimodal data on social media. *Computers and Electrical Engineering*, *120*, 109786., https://doi.org/10.1016/j.compeleceng.2024.109786 **[Paper Published - October 2024] [SCIE Indexed]**

Conference Publication

[1] Saraswat, P., & Beniwal, R. (2024, July). BERT-based RNN for Effective Detection of Depression with Severity Levels from Text Data. In *2024 IEEE Symposium on Wireless Technology & Applications (ISWTA)* (pp. 52-56). IEEE. **[Paper Published – July 2024] (Scopus-Indexed)**

[2] Saraswat, P., & Beniwal, R. BERT Based RNN for Effective Detection of Depression with Severity Levels from Text Data, IEEE Symposium on Wireless Technology and Applications. IEEE, 2024. **[Accepted and Presented] (Scopus-Indexed)**

Communicated Articles

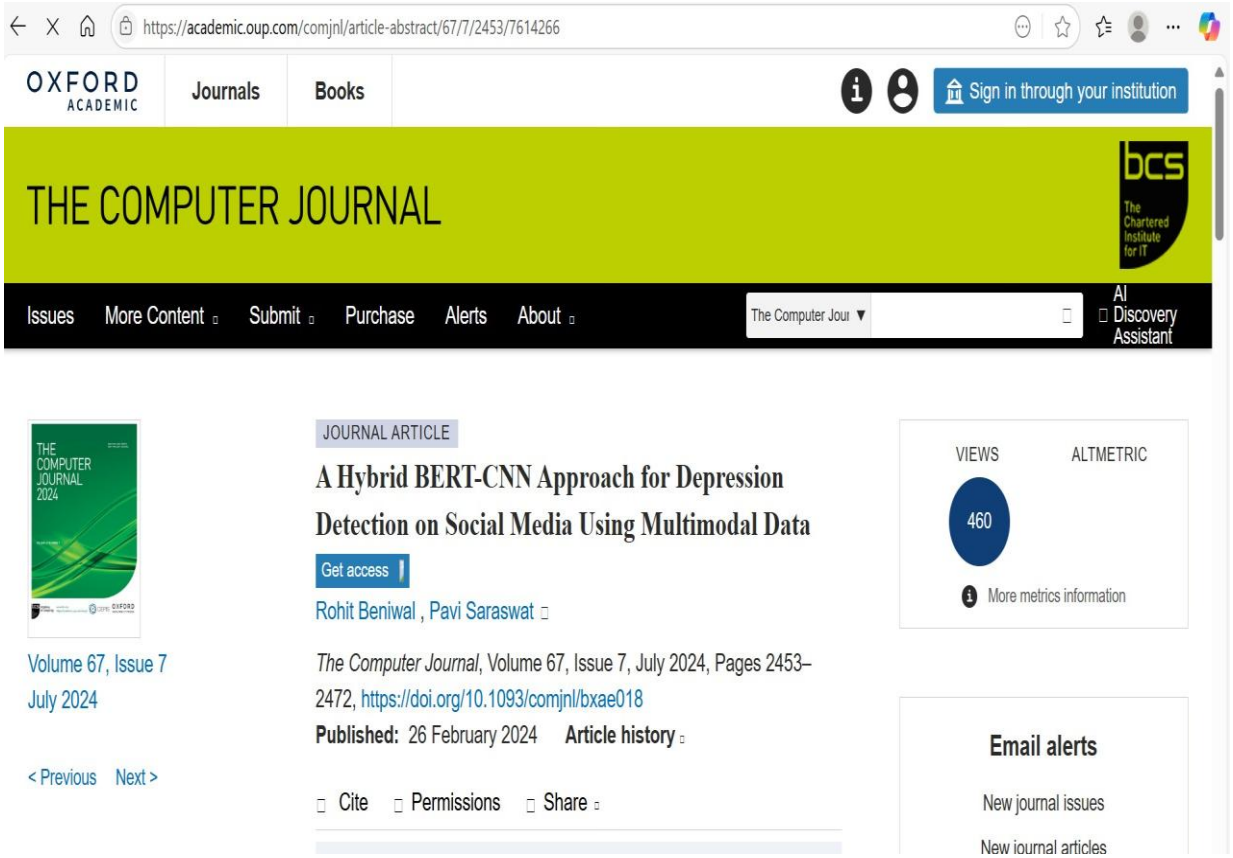
[1] Saraswat, P., & Beniwal, R. An Automated Hybrid Model for Depression Detection based on Vocal Features using Social Media Data Arabian Journal of Science and Engineering, Springer. ***(Under Review)***

[2] Saraswat, P., & Beniwal, R. Computer Assisted Optimised Hybrid Deep Ensemble Approach for Depression Analysis based on Facial Video Data, The Journal of Super Computing, Springer. ***(Under Review)***

[3] Saraswat, P., & Beniwal, R. DViTRoM: A Transformer-Based Multimodal Framework for Depression Detection Using Text and Visual Cues, Sādhanā, Springer ***(With Editor)***

SCIE JOURNALS PROOF

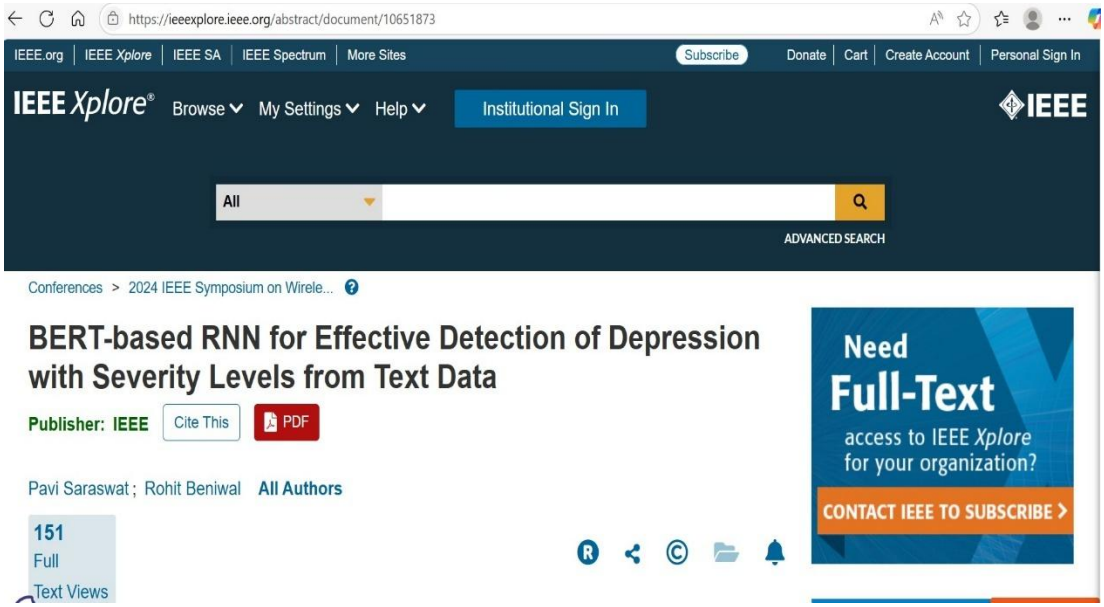
[1] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CNN approach for depression detection on social media using multimodal data. *The Computer Journal*, *67*(7), 2453-2472. https://doi.org/10.1093/comjnl/bxae018, **[Paper Published – January 2024]**



[2] Beniwal, R., & Saraswat, P. (2024). A hybrid BERT-CPSO model for multi-class depression detection using pure hindi and hinglish multimodal data on social media. *Computers and Electrical Engineering*, *120*, 109786., https://doi.org/10.1016/j.compeleceng.2024.109786 **[Paper Published - October 2024]**

CONFERENCE PUBLICATION(S) PROOF

[1] Saraswat, P., & Beniwal, R. BERT Based RNN for Effective Detection of Depression with Severity Levels from Text Data, IEEE Symposium on Wireless Technology and Applications (ISWTA-2024), Malaysia. **[Paper Published – July 2024]**



[2] Saraswat, P., & Beniwal, R. A Systematic Survey of Depression Detection: Modalities, Datasets, and Learning Techniques, 2nd International Conference on Computing and Machine Learning (CML 2025), Sikkim Manipal University, India. Springer. **[Accepted and Presented March 2025]**

Springer

2nd International Conference on Computing and Machine Learning (CML 2025)

organized by
Sikkim Manipal University, India

CERTIFICATE OF PRESENTATION

This is to certify that

Pavi Saraswat

has presented the paper titled **"A Systematic Survey of Depression Detection: Modalities, Datasets, and Learning Techniques"** authored by **Pavi Saraswat, Rohit Beniwal** in the 2nd International Conference on Computing and Machine Learning (CML 2025) held at Department of Computer Applications, Sikkim Manipal Institute of Technology, Sikkim Manipal University, India during March 22-23, 2025.

SCRS\CML2025\PC\39

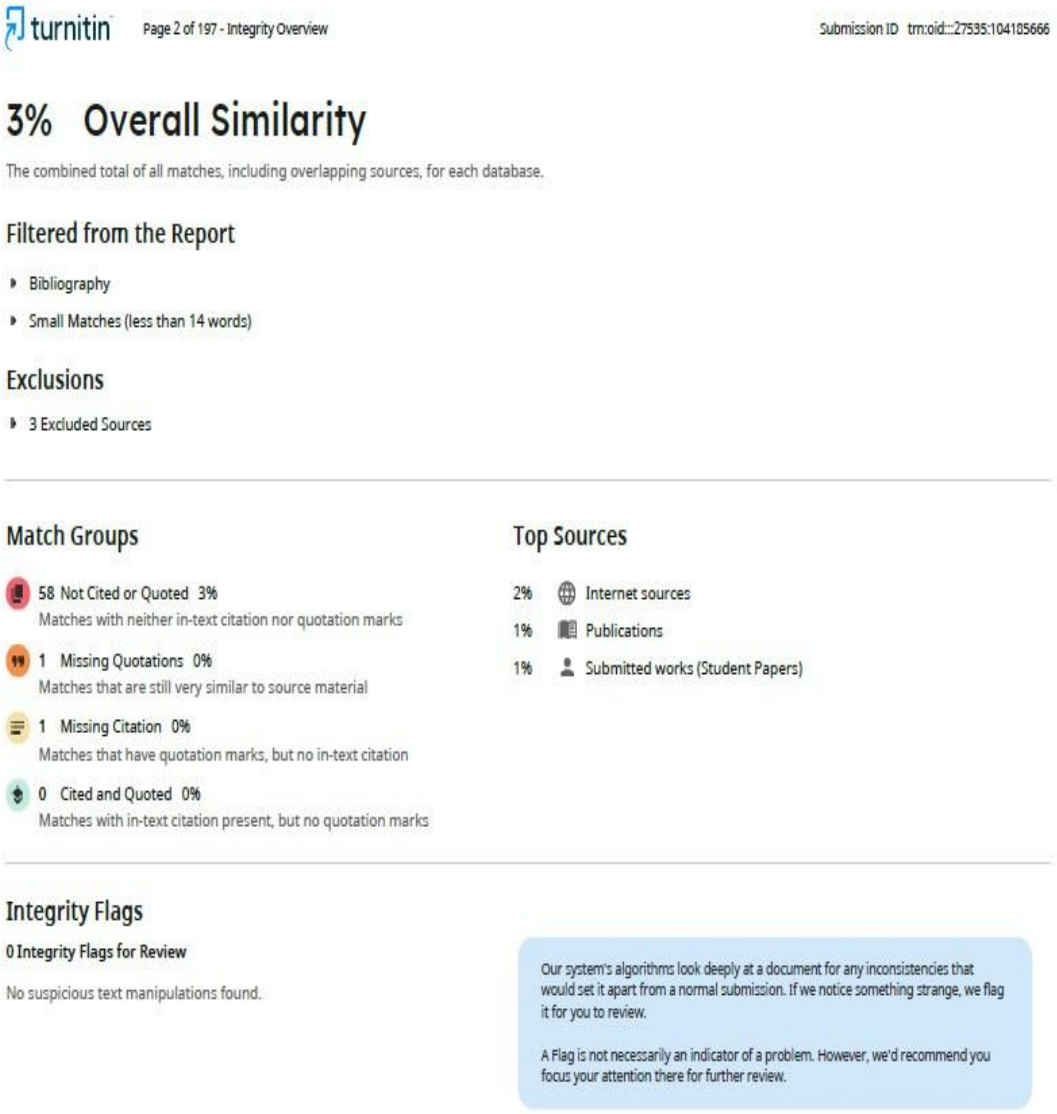
Prof. Shahid Hussain
General Chair

Dr. Samarjeet Borah
General Chair

https://www.scrs.in/conference/cml2025

Appendix B

Plagiarism Report


turnitin Page 2 of 197 - Integrity Overview Submission ID trn:oid:::27535:104185666

3% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 14 words)

Exclusions

- 3 Excluded Sources

Match Groups

58 Not Cited or Quoted 3%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 0%
Matches that are still very similar to source material

1 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
1% Publications
1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix C

Biography

Pavi Saraswat earned her bachelor's degree in information technology from the Dr. APJ Abdul Kalam University and her master's degree in computer science and engineering from the Amity University. She is presently at the Delhi Technological University pursuing a Ph.D. from Computer Science and Engineering department, Delhi, India.

Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Sentiment Analysis and Computer Vision with a passion for education and innovation. Her research centers on designing deep learning-based frameworks for depression detection. She also holds a patent in blockchain-based healthcare systems. She has published extensively in peer-reviewed international journals and conferences and actively contributes to the academic community.