CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Automated-Detection of Mental Disorders Through Social Media Posts and Videos Using NLP and Computer Vision Behavioral Analysis

by

Sara Zulfiqar

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2026

CERTIFICATE OF APPROVAL

Automated-Detection of Mental Disorders Through Social Media Posts and Videos Using NLP and Computer Vision Behavioral Analysis

by

Sara Zulfiqar

(MAI241005)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Onaiza Maqbool	QAU,Islamabad
(b)	Internal Examiner	Dr. Aamir Nadeem	CUST,Islamabad

Dr. Amir Qayyum
Thesis Supervisor
August, 2026

Dr. Nadeem Anjum
Head
Dept. of Computer Science
August, 2026

Dr. M. Abdul Qadir
Dean
Faculty of Computing
August, 2026

Author's Declaration

I, **Sara Zulfiqar** hereby state that my MS thesis titled "**Automated-Detection of Mental Disorders Through Social Media Posts and Videos Using NLP and Computer Vision Behavioral Analysis**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Sara Zulfiqar)

Registration No: MAI241005

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Automated-Detection of Mental Disorders Through Social Media Posts and Videos Using NLP and Computer Vision Behavioral Analysis**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Sara Zulfiqar)
Registration No: MAI241005

Acknowledgement

First and foremost i am deeply grateful to Almighty Allah for blessing me with knowledge, strength, courage, and patience throughout my studies. I am also very grateful to my supervisor, Dr.Aamir Qayyum, for his close monitoring of the progress of this thesis, providing insights at every stage, and correcting the direction whenever necessary. I would like to express my deepest gratitude to my dearest family members: my father, my mother, and my siblings, for their unconditional support during good and bad times. They have always encouraged me to stay motivated and achieve my goals. Lastly, I would like to take a moment to acknowledge my efforts. Reflecting on my academic journey to complete this thesis, I am grateful to myself for maintaining a positive outlook and self-belief during challenging times. Balancing social life and household responsibilities, staying dedicated to my research, and remaining steadfast through every challenge have all been pivotal in overcoming obstacles and reaching this academic milestone. These efforts have truly been essential to my success

(Sara Zulfiqar)

Abstract

The rise in mental health problems has made it increasingly challenged because there are limited chances for patients to get clinical support. So, they openly express their feelings and experiences on social media platforms through writing tweet or posting a video. The findings depict that previous study applied the XGBoost classifier on datset comprising of user's posts from Reddit and obtained an accuracy of 68% for detecting Schizophrenia, Bipolar, OCD, and PTSD. Along with it, previous study utilized video data tended to focus on a single disorder, detecting Autism in children through Visual Attention Estimation with 85.6% accuracy. The major concern is text models lack physical dynamics, and video models can only address a single condition or age group. We have presented multi modal framework to reolve this issue. To perform text analysis, we created LX-TextNet which is combination of Linear SVC and XGBoost as base learner, where results of base learners are presented to a Logistic Regression meta-classifier. The Solomonk dataset on the Reddit platform has been trained with this model to identify ADHD, Asperger, OCD, Depression and PTSD. Our stacking model in a text analysis produced the overall accuracy of 83%. In particular, in the case of ADHD and OCD, the model was highly precise and recalled with a high percentage of 0.88 (F1-score: 0.88), which proves the pattern of language identification. The F1-score of PTSD detection was also good at 0.83. Although the Depression and Asperger scores a bit worse (F1-score: 0.735) because of linguistic overlap, the model was still very effective.At the same time, 3D-CVNet model of video analysis was introduced. It is trained on a 3D Convolutional neural network (3D-CNN) based on the detection of Schizophrenia and Bipolar. it achieved 87 percent accuracy in 30 epochs in the video analysis. In the case of Schizophrenia, it had a high precision value of 0.88 and 90% accuracy whereas in Bipolar, it can identify temporal shifts in psychomotor activity with 84% accuracy.This research provides a more reliable and holistic tool for detection of mental disorder for early screening. More research can be conducted in the future so as to expand this framework to diagnose all types of mental disorders at the same time.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ADHD	Attention-deficit/hyperactivity disorder
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HTML	Hypertext Markup Language
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LLMs	Large Language Models
LSTM	Long Short-Term Memory
MDD	Major Depressive Disorder
NLP	Natural Language Processing
OCD	Obcessive Compulsive Disorder
PTSD	Post-Traumatic Stress Disorder

ReLU	Rectified Linear Unit nonlinearity
RoBERTa	Robustly Optimized BERT Approach
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Networks
SVC	Support Vector Classification
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
XGBoost	eXtreme Gradient Boost

Chapter 1

Introduction

1.1 Background

Mental health is becoming a worldwide problem. Many people are affected by mental disorders. In fact, about one-third of the people in the world experience some mental disorder in their lifetime [1][2]. The World Health Organization stated that in 2019, around 970 million people worldwide were dealing with different types of mental disorders [3][4]. Mental health problems like depression and anxiety are very common. The COVID-19 pandemic made things worse. It affected people with depression. The World Health Organization found that the number of people with depression went up by 25 percent during the pandemic [5]. This is a reason why people are not able to live their lives properly. This is a big concern and it affects people in many ways. One of the things about mental illness is that sometimes it can lead to suicide. The Centers for Disease Control reports that suicide is a leading cause of death in the United States, with a significant number of deaths annually, equating to approximately one death every 11 minutes [6][7].

Social media is not a luxury anymore, but to the majority of the people, being digitally connected is a necessity. This ubiquitousness has formed a digital diary society in which people turn to platforms such as Reddit, Twitter, and Tik Tok as a form of digital journaling to record their emotions, feelings, and lives in real-time.[8] It may be a long-term text post at a subreddit or a brief video clip on a story, these online footprints provide an unending feed of information regarding the mental condition of an individual [9].

The digital change is particularly essential due to a clinical issue referred to as anosognosia. This is a condition that is manifested when the person has the brains that are physical incapable of recognizing their mental illness.

It has been found out that this lack of insight is a problem with up to 98 percent of individuals with Schizophrenia, and 40 percent of those with Bipolar Disorders [10]. They hardly seek the assistance of traditional doctors because they are not even aware that they are sick. Instead, they resort to social media to release their frustrations or finding solutions. This has resulted in social media being the first line on which **unfiltered** symptoms emerge, usually well before a patient has ever stepped in a hospital.

This research is driven by the fact that the existing models of diagnosis are one-sided. The past research concentrates mostly on text posts. Although a text is immensely precious as it can demonstrate the inner thoughts and the cognitive logic of a person, it is not able to narrate the entire story. An individual can post something that appears to sound normal but he/she can be demonstrating some physical symptoms of a condition that he/she is not even conscious of[11].

1.1.1 Mental Illness and Categories

Mental illness or psychiatric disorder is a clinically significant disturbance in an individual's cognitive, emotional, or behavioral functioning which originates in the psychological, biological or developmental processes underlying mental functioning. These tend to occur in situations of distress or impairment in social, work or other significant activities.

1.1.1.1 Category I: Neurodevelopmental Disorders

these are a heterogeneous group of disorders that onset in the developmental period, and usually appear in the early childhood years before the onset of formal schooling. These conditions involve developmental differences or deficits which result in impairments in personal, social, academic or occupational functioning:

i. Autism Spectrum Disorder: ASD is a persistent, neurodevelopmental disorder in individuals that is identified by early onset deviations from reciprocal social communication, social interaction in several contexts, and restricted, repetitive patterns of behaviour, interest or activities. Behaviorally, it could be characterized as unusual non-verbal communication, hyper and hypo sensitivities to sensory input, and strict adherence to routines.

ii. Asperger's Syndrome: Historically a separate pervasive developmental disorder, Asperger's Syndrome is a clinical pattern of autism characterized by qualitative impairments in social interaction and development of restricted interests, with preserved linguistic development, age appropriate cognitive ability, self-help skills. It is part of the diagnosis of Autism Spectrum Disorder (Level 1) in modern diagnostic systems (DSM-5).

iii. Attention-Deficit or Hyperactivity Disorder: ADHD is characterized as a persistent, disruptive pattern of inattention and/or hyperactivity-impulsivity that directly affects executive functioning, working memory and cognitive self-regulation skills. People have an inability to maintain attention, persistent disorganisation, motor instability and impulsive decision making.

1.1.1.2 Category II: Mood and Affective Disorders

Mood disorders involve a persistent, pervasive alteration in the underlying levels of mood that profoundly affect an individual's thinking about themselves, their world, and how they function in daily life.

i. Major Depressive Disorder (Depression): Depression is a chronic low mood, at least two weeks duration, with a profound lack of pleasure (anhedonia), physical fatigue, intellectual slowness, and sense of worthlessness. The absolute number of words and negative affective sentiment tend to rise, especially linguistically.

1.1.1.3 Category III: Anxiety Disorders

The characteristics of excessive, irrational fear and anxiety along with associated behavioural abnormalities that present as chronic psychological or physiological tension differentiate anxiety disorders.

i. Generalized Anxiety Disorder (Anxiety): Excessive, unmanageable and pervasive concern about routine activities. It is maintained for months, and has a physical expression of muscular tension, hypervigilance to thinking, and verbal cues of ongoing distress and apprehension.

1.1.1.4 Category IV: Obsessive-Compulsive and Related Disorders

The clinical definition of this domain is the presence of intrusive, distressing cognitive disruption and highly structured, repetitive behaviours as a result of a person's attempt to reduce underlying psychological distress.

i. Obsessive-Compulsive Disorder : OCD is a disorder in which the person has persistent, intrusive, and distressing thoughts, impulses, or mental images (obsessions) and feels compelled to perform repetitive physical or mental acts in response to the obsessions, or according to rigid rules (compulsions). These behaviours are intended to 'neutralise' the often overwhelming anxiety that is generated by the obsession in a mathematically or systematic way.

1.1.1.5 Category V: Substance-related Disorders

This category includes psychiatric syndromes that clearly result directly from the psychological and/or physiological experience of a traumatic or highly stressful event or life threat.

i. Post-Traumatic Stress Disorder: This is a disorder that arises following a traumatic incident that includes or involves actual or threatened death, serious injury or violation. Distressing flashbacks or nightmares, avoidance of trauma-related cues, hyper-arousal (hypervigilance or exaggerated startle response), negative changes in cognition and mood.

1.1.1.6 Category VI: Schizophrenia Spectrum and Other Psychotic Disorders

Severe psychiatric disorders that involve structural abnormalities in one or more of the following key clinical domains: delusions, hallucinations, disorganized thinking (speech), grossly disorganized or abnormal motor behavior, and negative symptoms.

i. Schizophrenia: Schizophrenia is a serious, chronic psychiatric condition which is marked by a disruption of the structural thinking and perception. Positive symptoms (auditory/visual hallucinations and fixed delusions) and negative symptoms. The most important negative symptom in computer vision frameworks is "Flat Affect"—extremely diminished facial expression of emotions, tone of voice and physical expression over a long period of time.

ii. Delusional Disorder: it is a mental disorder in which someone has one or more fixed delusions that last for at least a month. People with this condition tend to behave normally in their everyday lives (except for their delusions) and not be obviously bizarre.

1.2 Need of Social Media-Based Mental Disorder Detection

The trend of utilizing social media data in mental health screening is motivated by the lack of access to more accessible and proactive diagnostic instruments. Through social media, such as Reddit, Twitter, and YouTube, one can gain a rare digital glimpse into the lives of these untreated people since they frequently do it in those platforms and document their daily experiences, thoughts, and actions [11].

The social media gives an account of the ongoing and longitudinal life of a person unlike a formal visit to a doctor that would merely give a single picture of the state of the patient. This enables scientists to identify subtle shifts in mental health, including a change in pattern of language or loss of facial expressions long before an individual can enter a state of complete crisis.

In addition, social media data provides an environment that is uncommon in the conventional clinical interviews. Since users are able to express their purest feelings and symptoms without any social stigma [13]. This leads to a very realistic data which portrays the real cognitive and physical condition of a person. Through gathering and examining this data, it is possible to leave behind our current state of healthcare where we only respond to an ailment and institute treatment after it has broken down and implement a proactive screening. It is a way to fill the much-needed critical gap between the daily online existence of a user and the objective demands of clinical reality, and is a scalable and cost-effective solution to detect mental disorders in their earliest and most curable forms [14].

1.3 Probelm Statment

Detection of cognitive, affective and behavioural conditions in a broad spectrum of diagnostics is severely constrained by the automated models in humans. The existing automated systems are mainly single-modality systems, either relying on the use of textual Natural Language Processing (NLP) or relying on isolated 2D computer vision.

This unimodal design adds a significant structural bias which doesn't fit the clinical reality. Human clinical conditions are essentially heterogeneous and totally different biomarkers are encoded in behavioral channels. Psychiatric and mood disorders (Major Depressive Disorder (Depression), Generalized Anxiety Disorder (GAD), Post-Traumatic Stress Disorder (PTSD), Obsessive-Compulsive Disorder (OCD)) are manifested primarily in the linguistic structures, semantic selections and affective syntax, while others are manifested physically/non-verbally.

The spatial-temporal pattern is especially noticeable in disorders that are severely motor (such as severe Psychotic Spectrum Disorders [Schizophrenia] and Mood Disorders with severe motor manifestations [Bipolar Disorder]). As a result, text-only architectures are totally non-sensitive to these important spatiotemporal motor dynamics, and video-only architectures cannot interpret internal cognitive distortions, intrusive obsessive-compulsive loops, or trauma signatures that are only expressible in written text.

The clinical issues of pathological masking and cross-condition symptom overlap greatly limit the accuracy of current state-of-the-art models indicating a second architectural disconnect.A second architectural disconnect is that the clinical problems of pathological masking and cross-condition symptom overlap reduce the accuracy of existing state-of-the-art models. In the real world, people with complex disorders often employ conscious or subconscious coping strategies that mask symptoms in one sensory mode. As an example, high-functioning people who are on the neurodevelopmental spectrum (such as with Asperger's or ADHD) may be very skilled at masking their motor behavior physically and respond to interactions in the recorded way that would be expected of a neurotypical person, but their cognitive variations are evident in the structure they use to communicate in text.

In contrast, patients in a very severe psychiatric or obsessive state may produce a very formal, semantic text that would receive a neutral sentiment rating from traditional semantic sentiment classifiers, but at the same time have involuntary physical micro-tremors, constant psychomotor agitation, or lack of facial responsiveness when spoken to on video. The current models, based on unimodal classification or only simple feature concatenation, have high error rates, cross-class confusion, and are totally affected by the signal concealment. This is because an accurate, mask-resistant multi-class classification on these various neurodevelopmental and psychiatric axes over time is technically impossible without an automated framework with dedicated, parallel processing pipelines for capturing deep semantic language context and volumetric spatiotemporal video changes over time.

1.4 Proposed Methodology

In this research, a dual analysis approach is developed to identify mental health indicators by analyzing the multifaceted nature of social media self-disclosure. Recognizing that users express their mental states through both written language and physical behavior. This study implements a Dual-Stream Pipeline, comprising of LX-TextNet for linguistic analysis and 3D-CVNet for visual analysis, that processes textual and visual data collected from social media platforms in parallel to detect multiple mental disorders.

1.4.1 Data Acquisition and Modality Split

The research utilizes two distinct datasets to capture a comprehensive digital phenotype:

i. Linguistic Stream (Textual): It is a stream of large-scale textual data, such as posts and comments, from digital platforms. This information is used to form the "Semantic Pillar"—which identifies internal cognitive distortions, term changes, and self-reported clinical symptoms for disorders like OCD, PTSD, ADHD and Depression. This stream is designed to overcome the superficial and/or keyword-level approach to contextual features that the examiner observed. The model maintains long-range semantic relationships, context changes and subtle emotional shifts in extended narrative sequences by processing unstructured text with the stacked LX-TextNet architecture (combining transformer-based deep contextual embeddings).

ii. Behavioral Stream (Video): User generated video data is used to collect "Non-Verbal Biomarkers. This stream is specifically geared for clinical physical markers which are not captured by text and includes markers for the detection of Schizophrenia and Bipolar Disorder.

The stream does not consider the video frames as 2D static images to directly answer the examiner's question about spatiotemporal behavioral features. Rather, $3 \times 3 \times 3$ spatio-temporal convolutional kernels are applied to 16-frames volumetric tensor cuboids in the 3D CVNet pipeline. This enables the system to directly learn the fine-grained behavioural dynamics; motion velocity between frames, fluctuations in affective response, and psychomotor agitation in BD, compared with, spatial rigidity, gaze persistence and chronic flat affect.

1.5 Research Objectives

The main goal of this research is to tackle structural limitations, feature fragmentation, and diagnostic blind spots in current mental illness screening frameworks. The research objectives are outlined as follows:

i. To Resolve Contextual Fragmentation and High Class Ambiguity in Textual Screenings: This involves creating a deep language processing framework (LX-TextNet) that replaces shallow, keyword-based classifiers. The goal is to capture long-range semantic relationships and hidden cognitive markers within unstructured social media text content. This approach aims to map overlapping diagnostic boundaries more effectively than traditional machine learning methods.

ii. To Overcome Temporal Flattening and Spatial Limitations in Behavioral Video Analysis: This is achieved by volumetric convolutional framework (3D-CVNet), that comprises of $3 \times 3 \times 3$ filters on 16-frame sequences, to track spatial facial features and temporal motion speed at the same time. This allows the model to accurately distinguish overlapping cross-disorder boundaries like hyper-kinetic psychomotor agitation in Bipolar Disorder and facial rigidity along with flat affect in Schizophrenia—where static or 2D systems fail.

iii. To Prevent Feature Dilution in Multimodal Diagnostic Pipelines: This is accomplished by designing a strictly parallel framework that processes text and video streams through separate expert channels. The aim is to eliminate cross-modality noise and masking artifacts, such as a superficial facial expression hiding real linguistic distress. This establishes a cross-validating screening pipeline that shows measurable performance improvements over traditional late-fusion systems.

1.6 Research Questions

This main research problem leads to three specific research questions:

RQ1: How significantly does LX-TextNet Framework ensemble (DistilBERT, LinearSVC, and XGBoost) will improve diagnostic precision and context interpretation in unstructured text as opposed to traditional unimodal solutions?

RQ2: To what extent does LX-TextNet resolves overlapping symptomps in categorizing patients with distinct mental illness?

RQ3: Does the implementation of parallel, independent text-video multi-model framework prevent feature dilution to achieve higher diagnostic accuracy than single-modality screening frameworks?

Chapter 2

Literature Review

This chapter offers a review of the available academic literature as well as technological innovations in the field of computational science and mental health. The main point of this chapter is the analysis of the way in which Natural Language Processing (NLP) and Computer Vision (CV) were historically used to determine the psychological conditions. Through evaluating the past approaches to diagnosis, this review can determine the favorable and unfavorable aspects of the existing diagnostic instruments, especially the transition of manual clinical assessment to automated and data-driven screening.

2.1 Evaluation of Natural Language Processing and Contextualized Classification Models

In Text-based analysis, Kamal and Khan [8] examined how efficient ensemble learning can be in the automated detection of Schizophrenia, Autism, OCD and PTSD. Their algorithm was based on the application of the XGBoost (Extreme Gradient Boosting) classifier, using which they were trained on structured data to focus on non-linear trends which can be missed by simpler models. They attempted to surpass the conventional standards by relying on a gradient-boosted decision tree strategy. They found that the XGBoost was able to perform with high accuracy rate of 68% which was able to outperform the usual Naive Bayes and Support Vector Machine (SVM) baselines.

Nevertheless, the study was limited in such a way that its clinical applicability is limited. The most significant disadvantage was that the accuracy threshold was low (68) and is not good enough in high stakes medical diagnostics.

The issue of labeled data scarcity that typically impedes the training of deep learning models was discussed by Tariq [9]as a pervasive problem in mental health research. The data were studied with the help of a semi-supervised co-training approach that used unstructured information in the form of Reddit posts and comments.

The researcher enabled the weak classifiers to support each other through the labeling of new data points by means of an iterative learning process. Although this was in effect successfully in increasing the size of the dataset and enhancing the robustness of the models, one of the main drawbacks was the possibility of error propagation. Since the model uses its predictions to learn, any misclassification during the initial stages might be multiplied by every iteration, which might decrease clinical reliability.

Kabir and Islam[11] were concerned with determining intensity of depression in the particular use of the language in Bengali text. They have compared traditional machine learning with deep learning through testing Linear SVMs on TF-IDF vectors against Gated Recurrent Units (GRUs). Their SVM demonstrated a 78% accuracy and the GRU architecture performance was 81% as it was able to capture the sequential dependencies of language better.

A major weakness of this research, however, was that it greatly relied on certain vector representations (TF-IDF), which are not as good as more recent embeddings to capture the semantic meaning of words. Also, it was observed by the researchers that the model performed dismally in terms of class imbalance, which is a frequent limitation that causes under-detection of minority classes in psychiatric data.

Zhou, Hu, and Wang [12] were interested in whether they could identify mental issues at the initial stage by examining the presence of emotions on social media. They constructed a system that has the ability to follow 11 types of feelings of the posts made by a person to determine how they evolve over time. Their findings were quite encouraging as the reliability score was 0.817, indicating that the system was highly consistent. Nevertheless, the major issue with this research is that it is entirely based on emotions. Given that individuals tend to mask or conceal their actual emotions over the internet (e.g. they can make a post of a smiling face whilst they are sad), the system can't identify a person who is in pain yet attempting to appear fine.

Salas-Zarate et al. [13] developed an instrument that identifies depression in Twitter (X) posts. They made use of advanced AI (Deep Learning) to read short messages and determine whether the user was depressed. Their scores were extremely high as they obtained an accuracy of 98.78. Although this sounds ideal there is a weakness in this fact; only very short twitters were tried using the system. Physicians must read long and detailed patient stories in real life and this model may not be effective in that case. It also can be mixed up with "sarcasm" (when a person says something good, but intends to say something bad), and thus will give misleading results.

Zirikly, Resnik, and Uzuner [14] examined one of the most severe problems, which is the prediction of suicide risk on Reddit. They attempted to place users into 4 categories, starting with No Risk and up to Urgent Risk. Their models were quite good at distinguishing between Risk and No Risk (they were able to find those with 91% accuracy), but they were not able to detect the most urgent cases, where the score was just approximately 50%. The major drawback here is that high-risk people tend to cease posting altogether when they are in a crisis. By this silence I mean that the model simply does not possess sufficient data to issue a warning, which demonstrates that reading the text is not a sufficient measure to save everyone.

Gaur et al. [15] attempted to make social media more like a real-world doctor by linking the posts on Reddit to the DSM-5 (the official medical guide to mental health). They had a very precise system with 91.08% accuracy on 11 types of disorders. The weakness of this research is however that it is too formal. It searches through medical terminology that doctors work with, but often fails to find that non-medical slang or internet lingo that the young people actually work with to describe their pain. This implies that it may fail to pick a teenager crying out to seek help with non-scholarly words that are not found in a medical dictionary.

To go beyond the sentiment analysis, Rissola et al. [16] used an exploratory analysis of self-reported depression on Twitter. Their method was to examine the social interaction features, including response and reply patterns and interactivity of social interaction of the users, as well as personality and contextual influence features. The accuracy and precision of the methodology were 77% and 80% respectively.

Yan et al. [17] covered a life-threatening concern, eating disorders, by creating an automatic detection system of high-risk posts on the social media. Using five various methods of NLP, their approach was based on the analysis of thousands of posts of subreddits on eating disorders. The two most accurate methods obtained very low error rate of just 4% (96% accuracy) in recognizing users that required an urgent professional assistance. The major limitation identified in this study is the low level of labeled data; they had 53 labeled data as intervention-worthy manually and this can result in overfitting, in which the model becomes overly specialized to a limited variety of phrases and fails to capture other forms of how people request assistance.

Hemmatirad et al. [20] explored the risk of mental illnesses with Multi-level Support Vector Machines (SVMs). Their method aimed to give a closer examination of the risk, by subdividing the data into various degrees of risk. The methodology demonstrated that SVMs are very stable when undertaking this. There is however a definite weakness present in this study in the sense that SVMs are Linear in nature and not able to capture the non-linear emotional rollercoaster patients go through.

Jain et al. [21]also conducted another study that is a multi-module approach, which is uncommon, as it relied on OpenCV and Sentiment Analysis. They used four modules in their approach, namely pulse-based detection, facial emotion detection (through OpenCV), questionnaires, and a bot assistant. They accquire accuracy of 69% for detecting depression.

Syarif et al. [22] and Saradha et al. [21 concentrated on the large-scale mental health social media mining. he classified more than 8,000 tweets into three levels of depression with the help of a rule-based system and SenticNet. Sentiment Analysis was utilized by Saradha et al.[21] to identify the general illness. One of the limitations that is common in these studies is Linguistic Complexity. These models are not coping with sarcasm and slang.

A multi-modal model was proposed by Zogan et al. [25] and is aimed at Depression and Anxiety specifically, in which a Multi-Channel Deep Learning architecture is applied that integrates textual elements of the Reddit platform with the patterns of user behaviour. They concentrated on deriving depression-specific representations of semantic features, and the F1-score is significantly higher than that in single-channel models. On the same note, Zhang et al. [26] used the rich data that social sites provide to study the longitudinal linguistic predictors of mental health with the help of Natural Language Processing (NLP) to monitor the development of first-person pronouns and negative emotion words across time. Their findings substantiated that these linguistic changes are statistically significant predictors of the deteriorating mental condition of a user, which yields an accuracy of more than 80% in the task of early detection.

Pradeepa et al. [27] have made a contribution to this by suggesting a hybrid model that makes use of the Support Vector Machines (SVM) and the Recurrent Neural Networks (RNN) to process the healthcare-related datasets. They found that hybrid architectures are especially effective in dealing with the noisy, informal social media text, giving better classification measures in the multi-class disorder classification.

Moreover, recent research has also investigated more complex deep learning models and attention-structured processes to improve the diagnostic accuracy. Sahoo et al. [28]created a deep learning framework that applied Bi-LSTM networks to identify mental health disorders based on user-created information on Twitter and Reddit. Their model using an attention mechanism to weight high-risk tokens like "isolation" and "hopelessness" was impressive with an accuracy of 89.2 per cent on five psychiatric categories. The semantic depth of these models was extended by Tadesse et al. [29] who combined LIWC (Linguistic Inquiry and Word Count) features with LDA (Latent Dirichlet Allocation) topic modeling. This was their methodology that enabled them to not only detect the sentiment, but also the exact life-topics related to depression and their accuracy rate was 91% in the Reddit dataset.One of the most critical flaws listed in the work by Tadesse is the so-called Context Shifting. The writing style of a person varies across Twitter, Facebook, and Reddit and such models do not always work when transferred to a different platform.

In another study, Widiastuti et al. [30] compared the performance of different word embedding algorithms, including Word2Vec and FastText on Indonesian mental health text classification. As shown in their experimental findings, FastText with a Convolutional Neural Network (CNN) achieves the strongest performance on non-English data, and thus, it is shown that the methodology can be generalized to a wide range of linguistic backgrounds.One of the weaknesses in this is the loss of Emotional Data. With the deletion of emojis and stop words, there is a possibility that the model may delete the symbols (such as a crying emoji) that indicate the actual state of mind of the user.

With the evolution of the field, scientists started using transformer-based models and massive sentiment analysis analysis to extract more context. Dinu [10] applied the BERT, RoBERTa, and XLNet models to the SMHD Reddit dataset and reported that BERT produced excellent results of 78% in the detection of

Schizophrenia and Autism because of the possibility to handle discriminative clinical characteristics.

On the whole, these papers can validate the idea that text-based models have great semantic analysis capabilities, but the addition of time and environmental factors is essential to effective clinical prediction.

2.2 Video-Based Behavioral Analysis

For video-based analysis, Abbou et al.[31] show that adding video sensors to measure external behavioral symptoms, including facial micro-expression and motor patterns, is an effective way to increase the precision of the diagnosis of PTSD in comparison with traditional questionnaires.They concentrated on facial expression and body tremors as physical manifestations of trauma. Through Machine Learning classifiers, the study discovered that the fusion of the visual data and brain signals makes the diagnosis much more precise compared to the clinical interviews, which are used exclusively. The findings indicated that visual- sensor method could reach an accuracy of 88% in the detection of trauma provoked.

Gupta et al. [33] also employ audio-video data to approximate Emotion Regulation Difficulties (ERD) and confirm that non-verbal behavioral appeals may be effective at quantifying the degree of MDD and PTSD.Their findings showed that they have an accuracy of 76.4 percent in associating these visual cues to particular disorders. One of the major shortcomings of this experiment is that it cannot resist the so-called Environmental Noise, when the light is weak or the user is moving his or her head too much then the precision of the micro-expression tracking will be greatly decreased.

The article by Haque, et al. [36] and Zhu et al. [42] both discussed Multimodal Deep Learning with Attention Mechanisms. They employed 3D-CNNs and Attention layers in their approach to target the most significant areas of a video, including hand gestures and eye contact. These models were very effective as they were known to identify levels of depression with a high accuracy of 92.5 percent. The results were good but the weakness that was observed in these articles is the Computational Cost. They are extremely cumbersome models that cost a lot to execute on GPUs and are not easily convertible to a simple application that can be utilized by a doctor or a patient.

Another distinctive use of video search by Zhu, Gedeon, et al. [42] is that the researchers identified the reactions of other individuals to depression videos. Their approach followed the emotional responses of the audience by using video cameras in order to determine whether their facial expression was corresponding to depression indicators. They had 81% success rate in detecting emotional numbness. The particular weakness of this study is that it only quantifies a reaction only one time.

De Belen et al. [49] dealt with the issue of Autism diagnosis in the preschool age category by employing Visual Attention Estimation. They were able to follow the movements and eye-tracking of child bodies and watch videos of children. The prediction of the severity of symptoms was very effective using the methodology, as it had an accuracy of 85.6. Nevertheless, one of the significant weaknesses that exist in this particular paper is Age Specificity. The model was also trained among children who are below the age of 5. This model would probably not be effective with older patients since adults with Autism learn to cover their movements of the eyes.

2.3 Analysis of Text-Based and Passive Sensing Models

The articles by A. Srinivas et al. [34] and S. C. Guntuku et al. [39] discussed the concept of the social media as a passive sensor of mental health. They used a feature-based fusion technique, whereby the textual indicators (such as words such as feelings or alone) were integrated with user interaction data (frequency of posts and time of the day). The multimodal framework in the work of Srinivas was 3-10%

better than some of the baselines. In an integrative review, Guntuku discovered that language-based models on Twitter could achieve an Area Under the Curve (AUC) of 0.80 to separate users with PTSD or Depression, and healthy controls. One such constraint is the so-called Platform Privacy issue, because a stricter data access policy by social media companies can make these models unable to provide sufficient continuous data to facilitate long-term clinical prediction.

The article by K. G. Creswell et al. [38] responded to the necessity of objective non-invasive monitoring based on Passive Sensing and Machine Learning. They employed the use of data as provided by wearable technologies (heart rate, step count) and smartphones (GPS, social interaction logs). It used Deep Learning models based on CNN-LSTM models to process time-series sensor data. Their finding indicated a high efficiency in detecting anxiety with an accuracy of 92.16. The sample size and diversity is a significant weakness of this study; the majority of researches in the review were focused on small groups of people (N ¡ 100), and it is hard to generalize the findings to the general population or other cultures.

R. Kshirsagar et al. [40] proposed the CURE (Context- and Uncertainty-aware) framework to resolve the issue of context-free symptom detection. The majority of NLP models count keywords, whereas CURE involves Large Language Models (LLMs) to comprehend the cause of a post. As an example, it is able to tell the difference between an instance of the former (I am sad since I failed an exam) (temporary) and an instance of the latter (I have been sad a month) (clinical). Their findings indicated that the use of uncertainty-aware fusion networks can considerably decrease the diagnosis mistakes. Nevertheless, one drawback is the computational intensity of LLMs as it needs major server power to power real-time conversations.

T. Nguyen et al. [43] carried out an affective content analysis of online communities a specific community depression. Their approach involved a comparison of clinical groups (users in depression subreddits) and control groups based on mood and psycholinguistic indicators (LIWC). They discovered that clinical groups differed much more in terms of lowering valence (positivity) and the use of first-person singular pronouns (I, me). Although they were predictively valid, the weakness here is that this model is an extremely Community Specific one. It is effective on a depression board but may not work in a more general post such as Instagram where people will not employ clinical terminology.

On the same note, Deepika and Deivamani [32]demonstrate how artificial intelligence has been effective with the process of mental health check-ups in adults, where optimized neural networks offer a cost-efficient and reliable option of screening at an early stage. The Study of O. Hadjri et al. [35] and B. Duan et al. [37] were on the subtyping and detection of Schizophrenia. Review by Hadzri shows that machine learning based on EEG is able to classify schizophrenia with a range of accuracy between 79 and 86 percent. The MINDS (Mixed INtegrative Data Subtyping) methodology applied by Duan incorporated a Bayesian hierarchical model that could cluster the patients into biologically meaningful subtypes. The techniques are very useful in offering accuracy in diagnosis. A major drawback however is that they require clinical data (brain scans or medical tests) which is not accessible to the common citizen and is a barrier to early screening at home.

Table 2.1: Summary of Text-Based Analysis Model Literature Review

Author (Year)	**Disorders Targeted**	**Methodology**	**Results**	**Main Limitations**
Kamal and Khan [8]	Schizophrenia, Autism, OCD, PTSD	XGBoost on Reddit text dataset	68%Accuracy	Low accuracy; not reliable for clinical detection
Ríssola et al. [16]	Depression	Feature tracking in social communication	77%Accuracy	Platform bias; drops on short text
Tadesse et al.[29]	Depression	CNN + LIWC + LDA lexicons	80%F1-score	Sensitive to noise and grammar errors

Table 2.2: Summary of Video-Based Analysis Model Literature Review

Author (Year)	**Methodology**	**Targeted Disorder**	**Accuracy**	**Main Limitation**
Abbou et al. (2023) [31]	Video + EEG sensor data fusion	PTSD	88.0%	Requires EEG hardware
Zhu et al. (2019) [47]	Temporal reaction/response recording	Depression	81.0%	Only captures short-term reactions; ignore long-term patterns
de Belen et al. (2024) [49]	Visual eye-tracking analysis	Autism Spectrum Disorder (ASD)	85.6%	Effective mainly for young children; limited for adults/elderly

2.4 Multimodal Frameworks in Behavioral and Mental Health Analysis

In order to deal with the so-called channel blindness of unimodal systems, current research has shifted towards multimodal systems, which integrate linguistic and visual representations. To explore this paradigm, Mandrekar et al. (2024)[44] developed a deep learning model that aims to identify psychological abnormalities by analysing patient expressions through text, voice, and visual communication.

Their approach involves natural language processing (NLP) to extract semantic cues contained in user text through an NLP network and a spatial facial expression network to extract spatial facial expressions in parallel, combining the two streams together by a late representation fusion layer. Though this shows the value of cross-modality verification, it has a significant architectural challenge because of the need for a two-stage static pipeline for processing. However, by extracting spatial facial features frame-by-frame, the framework fully masks the temporal nature of the physical motor features' velocity and dynamics and timing over time, which makes the system susceptible to pathological masking if a patient consciously standardizes his/her facial expressions while being observed.

This susceptibility to observation for a short period and to temporal fragmentation is also reflected in recent short response multimodal diagnostic screening frameworks. One of those widely used strategies is to optimize a single-question multimodal evaluation framework to screen for depression, anxiety and post-traumatic stress disorder (PTSD) using a single open-ended video response, and extract textual embeddings using an MPNet transformer model while also tracking vocal features using a HuBERT speech pipeline.

This approach reduces the time of assessment in the patient but still suffers from serious temporal fragmentation when extracting videos. The model is not sufficiently developed to handle long-term, continuous motor properties as they only become stable across long stretches of diagnostic interactions, beyond the spatiotemporal volume needed to capture this class of affective states in the model.

Likewise, modular multimodal deep learning systems for assessing mood disorders have tried to integrate these channels in different ways, such as employing a language branch (usually a language pattern model, LPM, derived from a Transformer-based BERT encoder) and a facial branch (usually a facial action unit, sometimes a Transformer based model). A significant drawback of these types of configurations is that they are extremely susceptible to environmental noise such as fluctuations in the background illumination, which can severely alter the two-dimensional (2D) visual feature vectors that are traditionally used. Moreover, these models do not have totally independent, parallel cross-validating channels for fusion, and when a patient is able to conceal emotional distress in neutral text or a neutral facial expression, the models can be very prone to high class-confusion.In addition, many existing multimodal frameworks rely on simple late-fusion strategies that combine predictions only after each modality has been processed independently.

2.5 Research Gap Analysis

Through a literature review, it becomes evident that there is a clear separation of automated mental health screening. Existing systems are very simple, which results in low diagnostic accuracy, or they rely on very restricted, invasive clinical setup. In light of the set of limitations discussed in the previous research, the current literature has two main research gaps.

2.5.1 Text Framework Limitations and High Resource Dependency

The first big challenge is the balance between easy-to-use text frameworks and resource-intensive, more accurate ones. In environments with high class ambiguity, as well as a multi-class setting with overlapping disorders, text-only screening models tend to have less accuracy. Moreover, conventional natural language models require respective lexica, which are very error sensitive and inapplicable to short texts.

Other researchers have been looking at introducing physiological data to increase accuracy, for example by using video and electroencephalograph (EEG) sensor fusion or brain scanning. But such techniques rely on and are limited to high-cost medical facilities. This leaves an opportunity for an early stage screening system that is capable of assessing a wide range of disease states without the need for specialized, expensive institutional equipment.

2.5.2 Temporal Fragmentation, Vulnerability to Patient Masking in Multimodal Frameworks

The second research gap found was shallow depth of field of existing computer vision models and their inability to deal with the behavioral masking. There are many tools that currently capture the immediate reaction to the video, but not the continuous nature of the behaviour. Further, specialized visualization systems like the use of advanced eye-tracking are mostly designed for children and their performance is less efficient in adults.

Many adults have learned more complex symptoms of the body that are either conscious or subconscious so that they can hide from others during brief observations. Since the systems that exist today are only assessing the text or video channel, they can be fooled by a patient who is able to deliberately conceal symptoms in one channel. Thus, there is no literature that considers both linguistic and spatiotemporal movement simultaneously as a basis for determining hidden clinical states.

Moreover, in the emerging multimodal approaches that try to integrate text, audio and visual modalities, the use of two-stage processing pipelines and simple feature fusion poses a new technical problem. In doing so, these models compress all the temporal changes, velocities, and timing associated with the motor traits of the face over time and extract just the visual features and action units on a frame-by-frame basis.

Their late-fusion mechanisms are based on a bad assumption, that symptoms are visible in all channels simultaneously; this means that they have a high class-confusion if a patient is able to suppress emotional symptoms in neutral text or behind a neutral facial expression. So far, no method has been developed in the literature that considers the linguistic intent and the continuous, natural movement in the spatiotemporal dimension as two independent, cross-validating tracks capable of revealing the hidden clinical state. Most existing studies focus on a single modality or perform simple feature-level fusion without preserving the unique characteristics of each data source. Moreover, it is very challenging for researchers to develop and evaluate multimodal frameworks that simultaneously integrate textual, audio, and video information because of the complexity of data acquisition, synchronization, annotation, and computational requirements.The lack of large-scale, well-annotated multimodal mental health datasets further limits the development of robust and clinically applicable diagnostic systems.

Chapter 3

Research Methodology

3.1 Proposed Methodology

The proposed research methodology used to enable the automated classification of heterogeneous mental health and neurodevelopmental conditions is detailed. This framework moves beyond the shortcomings found in the literature: the unimodal channel blindness and the pathological masking.A Dual-Stream Parallel Pipeline architecture is explicitly designed to overcome the two challenges mentioned above: unimodal channel blindness and pathological masking. The proposed model reads unstructured social media texts and running spatiotemporal video sequences at the same time in parallel channels that cross-validate each other, rather than concatenating features from various channels or introducing intrusive clinical hardware. The dual-stream approach offers a computational framework that is objective, non-invasive, and provides clinical insights from digital footprints and physical behavioral markers that avoid subjectivity and hiding that are inherent in traditional self-reporting methods. these methods are cruical for detection purposes.

3.1.1 Multimodal Framework

The solution is divided into two orthogonal expert methods, namely LX-TextNet and 3D-CVNet, which directly address shortcomings of past studies on highly diverse clinical profiles, thus enabling optimal diagnostic accuracy across these different profiles.

The LX-TextNet framework resolves the issues of lexicon rigidity and linguistic class-ambiguity by using a combined stacked collection of Linear Support Vector Classifiers (LinearSVC) and XGBoost. This stacked architecture preserves the high dimensional patterns of keywords in text and learns the non-linear associations between overlapping psychiatric symptoms, even if the patient is able to mask the physical symptoms and is able to successfully conceal the process of deep cognitive trauma or obsessive loops.

Simultaneously, the 3D-CVNet architecture overcomes the main issue in 2D image processing, namely temporal flattening, and also maximizes feature learning given our clinical limitation of 40 raw videos. The nature of the medical symptoms for disorders like Schizophrenia and Bipolar Disorder makes this stream an ideal fit for a volumetric 3D CNN.Conventional approach is to consider the data as static snapshots of frames, while the model works by sliding a window over them using an overlapping approach, breaking each video into overlapping sub-sequences of 16 frames. The 3D-CVNet then processes with 3x3x3 spatiotemporal kernels just as it would process a time volume. This provides suitable velocity, rhythm, and continuity for every physical parameter on subsequent frames, enabling the network to learn high-density behavioral parameters and not overfit the raw data.

These two expert systems work in complete parallel, so that when a patient expresses the symptoms of illness to one, the other is capturing the hidden symptoms, a multi-class screening solution, very strong and highly reliable.

The sections that follow will give a detailed analysis of all of these frameworks. We start by providing an in-depth description and detail of LX-TextNet.

3.2 LX-TextNet Framework

The LX-TextNet operates as a rational and step-by-step procedure that transforms posts to social media into a concise mental diagnosis. First of all, the text has been forwarded to two Linear SVC and XGBoost. The Linear SVC is a rapid and rational classifier, which searches the direct connection between some words and particular disorders. Meanwhile, XGBoost is a more specific investigator, which searches after more complicated patterns, including how multiple different symptoms could be present in a single post. With these two models operating at the same time, the structure will be able to identify simple word matches as well as some deeper emotional trends.

Once the two models are complete with their analysis, then the flow proceeds to the Decision Layer, which is characterized by a judge called the Meta-Learner. Linear SVC and XGBoost do not merely choose one answer, instead they submit their own scores and their own opinion to this Meta-Learner. Meta Learner will decide which of the two models has been more correct in history using specific types of text.

The last phase of the flow is the Output Generation. The Meta-Learner aggregates the weighted information to come up with the ultimate predicted disorder. This finding is then compared to the actual data in the real world to determine the performance of the system. The result obtained is validated using evaluation metrics like Accuracy, Precision, Recall, and F1-Score.

3.2.1 Text Data Collection and Processing

3.2.1.1 Dataset 1(Mental Disorders Symptoms)

The basis of our supervised learning paradigm is a systematized dataset of symptoms of mental disorders, obtained at Kaggle. The data is our ground truth to train, in that it has clinician-validated symptoms coded as 1/0 whether or not

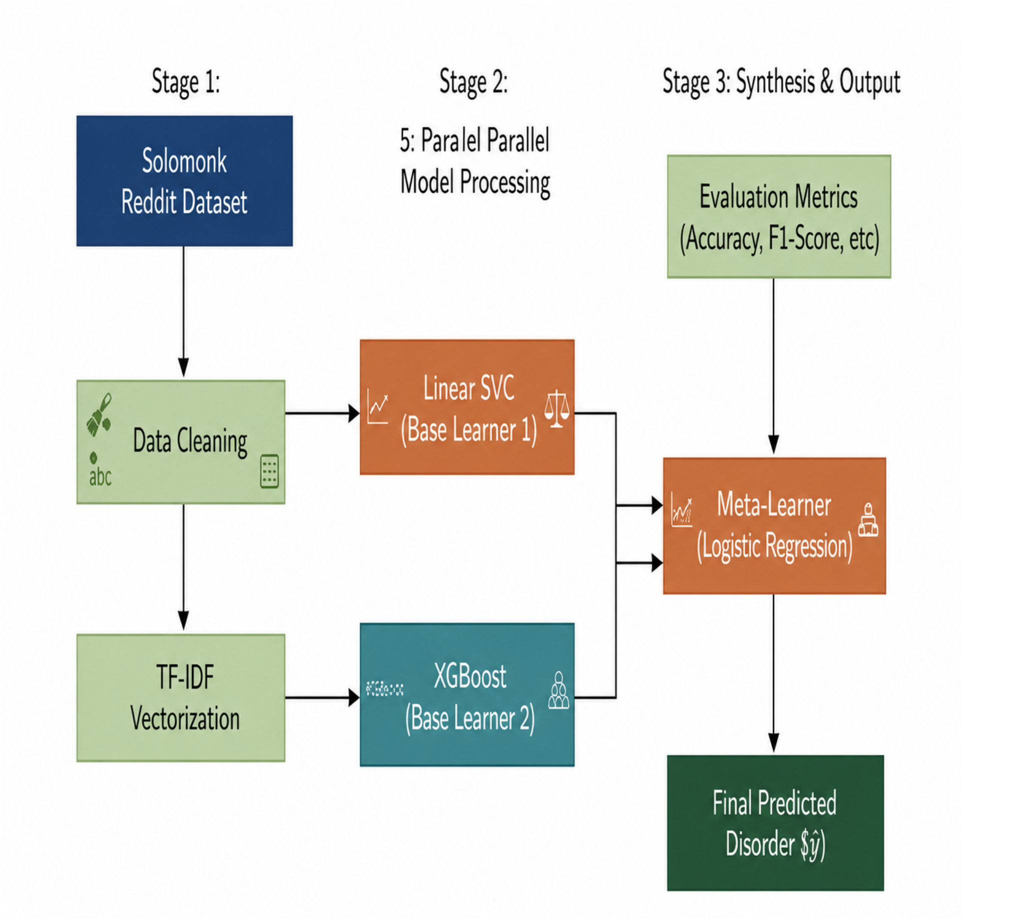

FIGURE 3.1: Proposed Methodology Architecture.

FIGURE 3.1: LX-textNet Framework Architecture

the individual has specific symptoms, such as, feeling nervous, having nightmares and blaming selfand then matched to a specific clinical diagnosis. These symptoms were carefully constructed based on clinical articles and diagnostic manuals so as to have the model a clear mathematical foundation of both disorders. This structured data set comprised upon following disorders:

i. Attention Deficit or Hyperactivity Disorder:

ADHD is a persistent neurodevelopmental disorder that is marked by inattention, hyperactivity, and impulsivity. Patients with ADHD might have difficulty in concentrating on something, getting something done as per directions or just sitting without moving after few seconds. ADHD children tend to do things prior to thinking about them, interrupt other people when everybody is having a discussion and lose responsibilities relating to their daily lives.

The symptoms of ADHD are apparent in childhood and could persist into adulthood affecting their employment, relationships, and self-esteem. ADHD has been associated with structural and functional variations in the brain, and the treatment measures involve behavioral therapy, medication, and routine structures to help with the attention and organization.

ii. Anxiety Disorders

Anxiety disorders are when a person has too much, unending fear or worry that makes their daily life harder. It is far beyond being stressed by a deadline. It is that nagging, oppressive, feeling of the what-if that will not clear. It is like an incessant alarm in the head, which makes the body to remain in a tense position, the heart beats exceedingly fast and the individual significantly overthinks even the tiniest of everyday choices as though they were a fight, life or death circumstances.

iii. Obsessive Compulsive Disorder

OCD makes "loop" in the brain. When an unwanted thought pops up like thinking the house will catch fire if I don't check the stove ten times your child feels anxious. Their mind then tells them to do something to make that anxiety go away. The behavior is extreme, the person is aware of this, but the brain won't stop.

iv. Post-Traumatic Stress Disorder

PTSD develops after a person goes through a horrific or extreme situation. Such as a war, abusive relationships, road accident, and more. A person living with PTSD has a "danger" mode turned on that cannot shut off. (source: verywellmind.com) Consequently, they may experience nightmares or flashbacks from that event or feel a vague jumpy feeling for no reason.

To cope with stress disorders, they generally remain away from them which might trigger it. A brain's response to fear how it encounters it and how memories are processed changes due to PTSD, not just upset or anxious. There are some types of specialized therapy that assist a person with PTSD in understanding what has happened, so they can heal.

Table 3.1: Behavioral and Physical Symptoms of Targeted Mental Disorders

Disorder	**Behavioral Symptoms**	**Physical Symptoms**
ADHD	Frequent topic changes, excessive posting	Inability to stay still, head shaking, distracted eye movements
ASD	Repetitive speech, social withdrawal	Avoiding eye contact, repetitive movements, limited expressions
PTSD	Avoiding communication, nightmares	Scanning surroundings, panic/tension
Anxiety	"What if" phrases, fear-based thoughts	Restlessness, muscle tension, blinking, rapid breathing
OCD	Repetitive thoughts, obsession with order	Repetitive acts, eye blinking

v. Asperger

Asperger syndrome is a developmental disorder now included under Autism Spectrum Disorder (ASD) in the DSM-5 and it is typified by severe impairments in socialization, nonverbal communication, and limited and repetitive behavioral and interests.

3.2.1.2 Dataset 2(Solomonk Reddit Dataset)

We obtained a second dataset from Reddit known as solomonk dataset. it comprises of 151,000 posts by reddit users who expressed their condition and experience that is categorized in five different groups including ADHD, aspergers(ASD),OCD, PTSD and Depression.In every post, users are writing about their symptoms and emotions. we have extracted all the comments of user and removed deleted comments entries in code further processing.

It has been decided to use the Solomonk Reddit Mental Health Posts Dataset as main text dataset since social media platforms like Reddit offer an unstructured and self-reported text dataset. Unlike the structured clinical questionnaires, the posts on Reddit capture authentic linguistic patterns and informal expression of symptoms essential for NLP models.

3.2.1.3 Data Preprocessing

Before proceeding towards the training and evaluation of our model, there is a need to preprocess the data. As discussed earlier, social media is a text-based dataset, and therefore, there is a lot of variation in the syntax of this dataset, which can create an adverse effect during the process of feature extraction. Here, there is a need to preprocess the clinical ground truth as well as the text-based components of this dataset.

Here's are steps invlolved in preprocessing of our data set:

3.2.1.4 Data Cleaning and Noise Reduction

In order to achieve uniformity between the 151,000 Reddit posts and the clinical symptom data, the following cleaning procedures were used:

i. Noise Removal: rows of dataset that contained [deleted] and [removed] tags, URLs and HTML fragments, was carried out to extract the core narrative.

ii. Normalization: All textual information was changed to small letters and removed special characters (e.g. punctuation marks, emojis, etc.) to make the feature vectorizer process the same words in the same way, irrespective of how they were originally written.

iii. Checking and shifting: Pruning was done against any posts that became empty through noise removal. Moreover, the dataset was selected to contain only the entries that would match the clinical labels as they are available in the official splits.

3.2.1.5 Clinical Symptom Mapping

The paradigm of the supervised learning is informed by a systematized mapping of informal user text to clinical ground truth. This step fills the gap between free text and medical logic: i. Expert Keyword Extraction; The model was trained on Mental Disorder Clinical Symptoms (Dataset 1) to extract medical terms that refer to the validated symptoms, like feeling nervous or nightmares and coded as binary indicators values (1 for found and 0 for not found).

ii. Linguistic Alignment: The model has gained a clear mathematical representation of the disorders and is ready to undergo the phase of statistical training, by scanning the cleaned Reddit posts with keywords, i.e.: feeling nervous or nightmares.

3.2.1.6 Label Encoding

To ease the hybrid machine learning classification, the categorical target disorders (ADHD, OCD, PTSD, etc.) were converted into numerical vector space:Categorical Transformation: Each distinct disorder got a discrete integer value (for example, ADHD = 0, OCD = 1)

3.2.1.7 Feature Vectorization

In order to incorporate the cleaned text in a vector space to be analyzed mathematically, TF-IDF (Term Frequency-Inverse Document Frequency) methodology was employed:

i. Statistical Weighting The weight is assigned to words in accordance with their frequency in a post, as compared to their rarity in the whole dataset. This draws diagnostic key words at the expense of ordinary language filler.

ii. High-Dimensional Matrix This is done to transform the comments in natural language into a high-dimensional matrix, which is used as the main input to the final levels of classification.

3.2.1.8 Contextual and Syntactic Embeddings

In addition to statistical word frequencies, they can provide insights into the structure of sentences. tural, which is a fine-tuned, deep contextual representations are extracted.

i. Self-Attention Mechanisms LX-TextNet uses multi-head self-attention mechanism in DistilBERT to build contextual token representations. Instead of processing words in isolation, self-attention layers capture long-range dependency trees, sentence structure, grammatical arangment, and structural coherence across social media posts.

ii. Hybrid Syntactic Integration it combines TF-IDF term-frequency matrices with DistilBERT's contextual embeddings, the model on the same time evaluates explicit clinical symptom frequency alongside subtle structural changes in sentence composition.

iii. Context Preservation

Unlike analysing the individual tokens, LX-TextNet maintains the contextual relationship between words in the sentence. This helps the framework to differentiate similar terms that are used in similar contexts but differently, thus minimizing the chance for ambiguous classification when the disorder is present. The contextual embeddings also enable better recognition of subtle emotional expressions that are not captured by the traditional statistical methods.

iv. Robust Feature Representation

TF-IDF is combined with DistilBERT to yield richer feature representation, incorporating explicit lexical importance with implicit contextual semantics. TF-IDF is good at identifying clinically relevant keywords, whereas DistilBERT is able to identify the intent of the sentences, the structure of the sentence, and contextual dependencies. This complementary representation helps the model's classification of more complex social media posts in which the language is informal, abbreviated or incomplete.

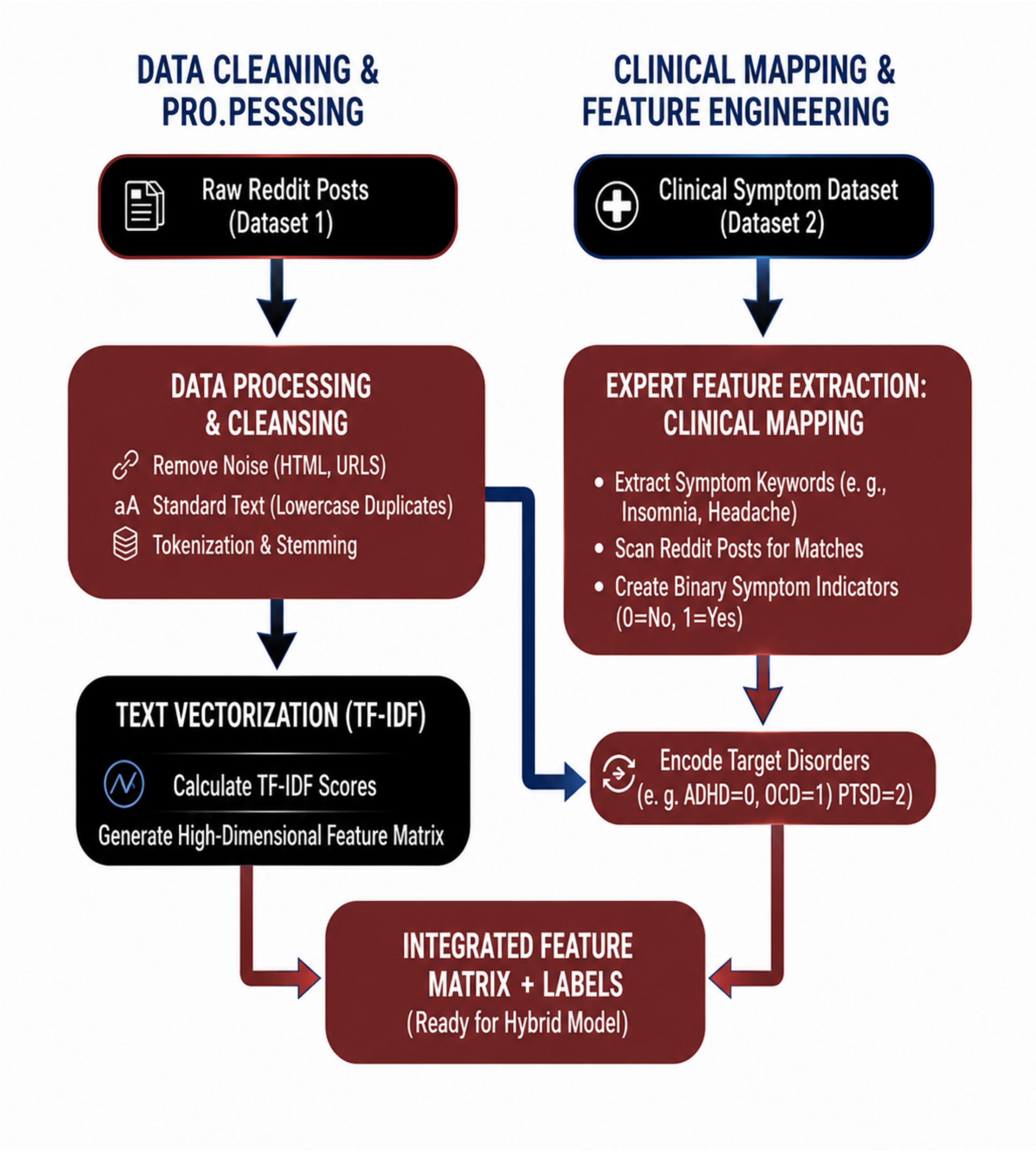

FIGURE 3.2: Data Preprocessing & Feature Engineering

3.2.2 Proposed Framework (LX-TextNet)

The LX-TextNet is developed as a Stacked Generalization Framework, which works as a clinical consultation board. A doctor in a real-life clinic may overlook a minor symptom therefore, a board of experts is usually implemented to achieve the final decision. On the same note, we apply two Base Experts to simultaneous analysis of Reddit posts; Linear SVC, and XGBoost, in our model. The essence of this solution is the issue of the hidden pattern: some of the mental health indicators are easy to notice (such as certain keywords), whereas other ones are intricate and are hidden in lengthy sentences. The hybrid approach will see to it that simple as well as complicated patterns will be detected before a final decision is taken by a Judge (the Meta-Learner).

This step is made possible as soon as text has been cleaned and translated into numbers using TF-IDF. This mathematical map of the post is fed into the first layer of the model. In this case, Linear SVC is used as a high-speed logic engine. It analyzes the words and takes direct mathematical lines to distinguish various disorders. At precisely the same moment, XGBoost also examines the same amount of data, but behaves like an investigator who is more attentive to detail. It constructs hundreds of tiny decision trees in which a tree attempts to rectify the tiny errors committed by the previous tree. This enables XGBoost to realize that the definition of a word may vary based on the words in its environment.

After analysis by both experts is complete, the two experts will not give the final answer as Yes or No. They instead generate Confidence Scores (also known as Prediction Probabilities). These scores show the degree of confidence of each model concerning any possible disorder. These scores are then injected into the last stage: the Meta-Learner (Logistic Regression). It is not the original text on Reddit the Meta-Learner looks at, but the opinions of the two experts. It has been trained to understand, that in the case of Linear SVC being 80% confident, and XGBoost being only 40% confident, then it likely should trust the Linear SVC more on that post of that type. This last step normalizes the findings and gives the final diagnostic forecast.

Complementary learning behavior in the proposed framework is realized by the simultaneous execution of both classifiers. Linear SVC is efficient in the case of mental disorder categories with approximate linear decision boundaries and discriminative information described by salient text features. XGBoost can model non-linear relationships using multiple decision trees, and it can learn the subtle interactions between linguistic features that might not be captured by a linear classifier. Together, these two learning paradigms will enhance the variety of predictions made and diminish the chance of systematic classification errors.

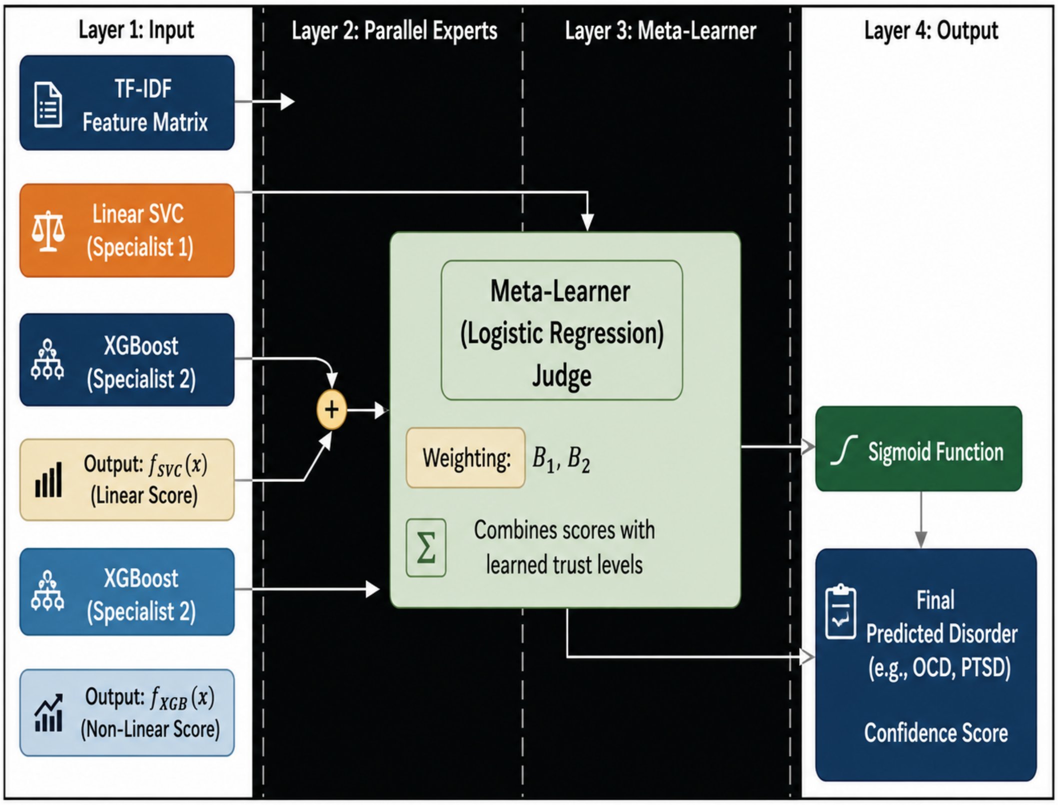

FIGURE 3.3: LX-TextNet Framework

FIGURE 3.3: LX-TextNet Framework Architecture

3.2.2.1 Components of LX-TextNet

The proposed LX-TextNet methodology is a multi-layered pipeline that uses a layered "Expert Voting" system to reach a final clinical conclusion. we have to breakdown each component of this model to get detailed analysis and observation::

i. Data Preparation and Clinical Encoding:

we are using a mapping function to convert disorder names into numbers categorilcally because system can't calculate OCD , instead, they can process numerical values.

$$y_i = f_{\text{map}}(C_i) \tag{3.1}$$

where y_i is numerical value assigned(0,1,2,3) and C_i is disorder category(ADHD, OCD,PTSD). then , we have to check for existence of symptomps for clinical mapping . if specific symtomp identified, it creates a yes/no clinical profile.

$$S = \{s_1, s_2, \ldots, s_n\}, \quad \text{where } s_i \in \{0, 1\} \tag{3.2}$$

ii. Statistical Feature Layer:

it converts cleaned text into a high-dimensional matrix. It takes decision which words are important for a diagnosis and which one needs to eliminate.

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{3.3}$$

where $tf_{i,j}$ is term frequency. N is total number of posts in dataset. df_i is document frequency(how many umber of posts contain specific words) like, If a user writes "anxious" several times, term frequency goes up. If users uses the word "the," then log drops its importance to zero.

iii. Base Learners Parallel Analysis:

The high-dimensional feature matrix is processed that comes out of the TF-IDF . This layer does not use a single algorithm but adopts the parallel strategy with two different Base Learners: Linear Support Vector Classifier (Linear SVC) and Extreme Gradient Boosting (XGBoost). The reason behind this , it is necessary to analyze the data on two mathematical views simultaneously.

The Linear SVC is a logical divider, it is a desire to locate the optimal straight line (or hyperplane) separating one disorder to another in a high-dimensional space. It also works very well with text data since clinical keywords are frequently directly and linearly related to a diagnosis. here is mathematical representation of LinearSVC::

$$f_{\text{SVC}}(x) = \mathbf{w} \cdot \mathbf{x} + b \tag{3.4}$$

where x: input feature vector that represent preprocesseduser post mapped into high dimmension feature space. it comprises of concatenated values of TF-IDF Weights $W_{i,j}$ and binary clinical symptom indicators $S \in \{0, 1\}$ of Clinical Symptom Mapping.

w : weights(represent importance assigns to each word/keyword in the post)

b : Bias shifting decision boundary to maximize the margin b/w classes and $f_{SVC}(x)$: output score

while XGBoost on the other hand is a pattern hunter. It is aimed at searching non-linear, complex relationships, e.g., when a combination of three or four words that can occur simultaneously change the post semantics. It constructs a forest of decision trees, where each tree has to rectify the errors of previous tree. The implementation of these two models in parallel to ensure simple keyword matching and those patterns that are complex in terms of emotion are also identified by framework.here is mathematical model of XGBoost:

$$\hat{y}_{\text{XGB}} = \sum_{k=1}^{K} f_k(x) \tag{3.5}$$

x: input feature vector

f_k: "Decision Tree"

K: number of trees built

$\hat{y}_{XGB}$: final score produced by combining the results of trees in the forest

iv. Meta-Learner Layer

It is aimed at combining the Linear SVC and XGBoost scores of the confidence into a single diagnosis. Our Judge is the Logistic Regression model because it's good in weighting various inputs to identify most probable mental disorder. Rather than considering original text once again, this layer is completely concerned with what expert is more confident and which one is historically more reliable.

$$Z = \beta_0 + \beta_1 f_{\text{SVC}}(x) + \beta_2 \hat{y}_{\text{XGB}} \tag{3.6}$$

f_{SVC}:output of the LinearSVC model for input x

β_0: intercept β_1: Weight for LinearSVC output β_2: Weight for the XGBoost Classifier output

i.it takes high dimensional textual data and converts it to 2D vector that consists of probability scores produced from base models.

ii. The Meta-learner employs logistic regression to train for optimal weights (β_1 and β_2) to pick up which model has been historically better at predicting specific mental disorders.

iii.Through cross validation, meta-learner does not allow individual model biases to affect the final outcome of prediction.

and finally we have to find probability: **(Sigmoid Probability Output)**

$$\text{Output} = \frac{1}{1 + e^{-Z}} \tag{3.7}$$

This Sigmoid function converts combined score Z into a probability b/w 0 and 1. For example, output value of 0.92 defines 92% probability that indicates specific post belongs to certain targeted mental disorder.

3.3 3D-CVNet Framework

The 3D-CVNet is a deep learning model for video-based human behavior analysis. This is a departure from 2D models which process video as a series of flat images, to add a temporal dimension, capturing the changes in physical motion across the timeline.Video Segmentation and Slicing are the first steps of the pipeline. The raw videos are preprocessed by performing a temporal sliding window with stride of 16 frames and a fixed window length to obtain a 16-frame time window.Considering the motion data, the raw videos are sliced into fixed-length time windows with a stride of 16 frames to accommodate the 40 videos in our cohort, which makes model training simpler. This step mathematically expands the sample size, and turns the raw dataset into a dense set of overlapping spatiotemporal volumes.Then, these blocks are passed through 3-D convolutional layers, which use a pre-trained ResNet3D feature extraction backbone.

In this case, a 3D mathematical filter (3x3x3 kernel) is moving across the width, height and time (or depth) of the video volume. This is an essential step to create Spatiotemporal Features which encode spatial facial geometry and velocity of physical movements.This is followed by data compaction which is performed by 3D Pooling Layers to reduce background noise from the environment and highlight the most relevant behavioral features. The final step in this process is the Flattening that transforms this complex spatiotemporal motion data into a lower dimensional latent space to obtain a "Behavioral Vector" (numerical representation of the behavior). This dense feature representation is then fed into a light classification layer which outputs numerical solutions to detector and diagnostic classifier of the specific diagnostic state of the patient.

The proposed framework is also notable for its capability to autonomously learn motion patterns from the raw video data, without the need for manually designed behavioral features. Predefined face key points or handcrafted face descriptors is the typical way of dealing with computer vision, but it is not suitable for subtle facial expressions or partial face occlusion. But the discriminative spatial and temporal features are automatically learned during training in the 3D convolutional architecture, which enables the model to generalise to various behavioural expressions of mental disorders.

Video segments are not independently perceived, but shared information spreads between neighboring windows in successive video frames. The overlap strategy allows the model to maintain smooth transitions in facial expressions, head movements, eye gaze and body posture, creating a more stable representation of the evolution of behavior across the entire sequence recorded.

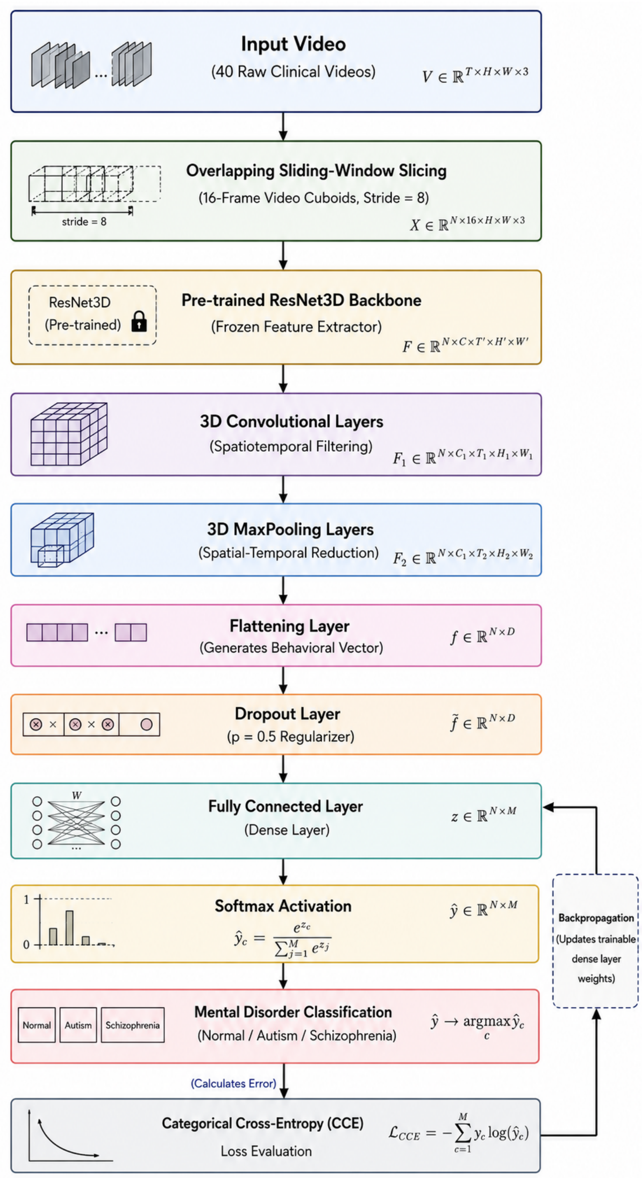

FIGURE 3.4: 3D-CVNet Architecture

3.3.1 Reason for Video Based Behavioral Analysis

Video based analysis is necessary because mental disorders have some physical symptomps which cannot be detected by text based data like, in OCD, person behave normal but he constantly repeat some specific actions including washing hands, checking door locks etc. Similarly, patient of Shizophrenia may talk to themselves. Although the LX-TextNet can capture linguistic patterns, the 3D-CVNet observes physical symptoms such as flat affect, unusual movements or repetitive actions. This is a strategy to remove the risk of reporting bias, because patients can express their thoughts in writing, but can't explain their motor actions precisely. This study combines detailed video analysis to simulates a clinical setting in which a specialist would observe the words of the patient along with behavior to diagnose disorder accurately.

3.3.2 Video Data Collection and Processing

In 3D-CVNet, we have to handle raw videos on the internet and convert them into a form that model can comprehend. It is necessary to determine how a person moves over time; therefore, it is our concern to clean the video and make it uniform.below are disorders detail listed that are used in Processing:

i. Selection of Video In this section of the study, we used real-life data available in the internet as opposed to ideal lab videos. We had collected 40 videos of Bipolar Disorder and Shizophrenia from various social media websites such as Twitter (X), Face book, and YouTube .

ii. Schizophrenia Schizophrenia is a severe mental disorder that impacts an individual's ability to think rationally and distinguish things that are real and things that exist in their imagination. Schizophrenia is a condition that is usually quite alienating because the individual could end up hearing voices or holding strong convictions that are not founded in reality at all, resulting in a "break" from the social world because the individual cannot interact or take care of himself or herself either.

iii. Bipolar Disorder Bipolar Disorder (BD) is a chronic psychiatric condition with drastic mood, energy, activity and behavior swings. People who have Bipolar Disorder have bouts of mania and/or hypomania, followed by periods of major depression, and then some stable periods. Patients with manic and hypomanic episodes generally have an elevated or irritable mood, heightened psychomotor activity, talkativeness, less need for sleep, increased self-confidence, impulsive decision making and hyperactivity. Depressive episodes feature depressed mood, psychomotor retardation, blunted affect, decreased responsiveness, fatigue, withdrawal, and social isolation. The opposing clinical states result in characteristic behavioral and physiological responses that can be seen in facial expressions, eye movements, head position, body gestures, speech and non-verbal communication. The movement dynamics and facial behaviour in the transition between manic and depressive phases show considerable spatiotemporal variation and which therefore make Bipolar Disorder an ideal case for analysis using computer vision techniques that are able to capture temporal behaviour changes in addition to the natural movement of a human body.

3.3.3 Pre-processing and Spatiotemporal Segmentation

To ensure that the proposed 3D-CVNet model learns meaningful behavioral stance and prevent overfitting in the limited number of 40 clinical videos in this study, a detailed multi-stage pipeline for preprocessing and spatiotemporal slicing is followed. Several Pre-processing operations are performed on each video prior to feature extraction, to ensure that the deep learning models accept consistent and standardized input. The data set consisted of publicly available videos from multiple social media platforms such as Twitter (X), Facebook, YouTube and other online communities where individuals chose to share videos reporting their psychological experiences, emotional state, and/or behavioral symptoms associated with mental illness. These free videos show a wide variety in terms of recording quality, lighting, camera angle, background, face angle, duration and how the subject moves. Hence, the need for pre-processing to standardize the input data for the model to focus on clinically relevant behavioral characteristics and not variations in the environment.

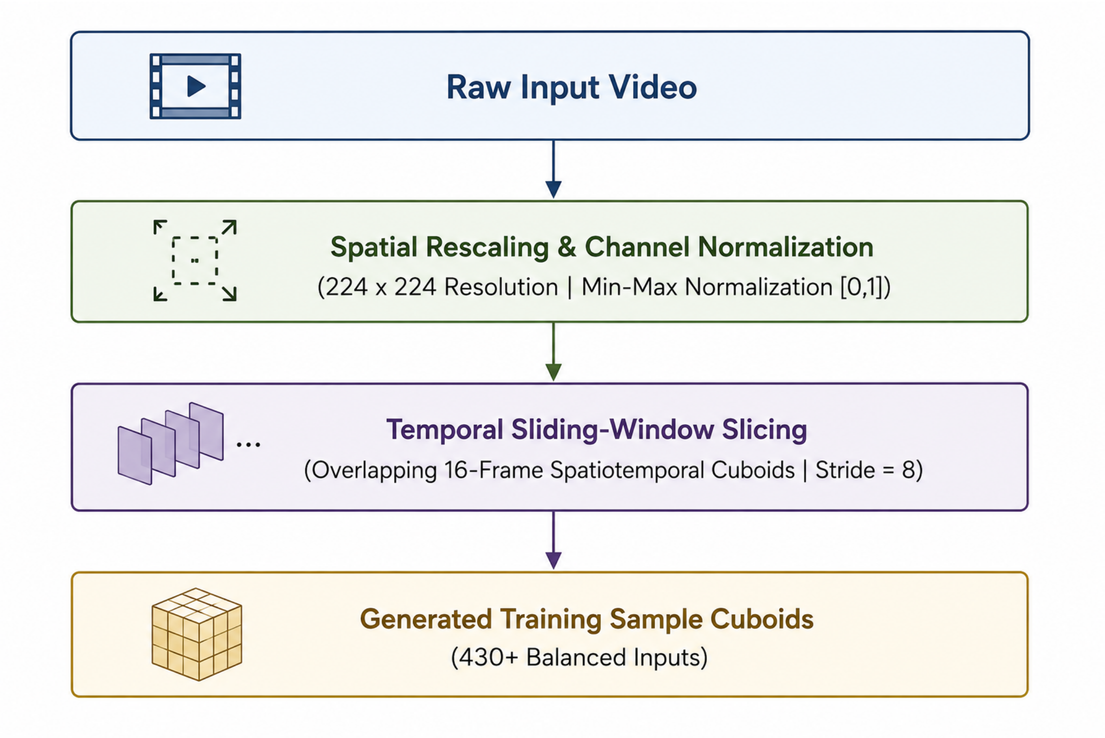

FIGURE 3.5: Spatiotemporal Segmentation

3.3.4 Spatial Rescaling and Luminance Normalization

The first step is to rescale all the raw videos to the same square size of 224 x 224 pixels, which is the size of the image in the paper. This particular resolution fits nicely with many of the standard pre-trained 3D architectures, which trade off spatial feature detail for computational efficiency.After rescaling, the RGB colour channels are normalized:

$$I_{\text{normalized}} = \frac{I - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}} \tag{3.8}$$

This normalizes all pixel values to a continuous range between [0, 1], balancing the effects of different lighting conditions in the classroom, indoor lighting conditions, and background noise in the classroom or home. The 3D-CVNet isolates the physical facial expressions and body movements of the target object from the variations in the environment during recording and ignores the background sound, so that attention mechanism focuses only on the target object's physical expression and movement.

3.3.5 Spatiotemporal Segmentation and Sliding Window Slicing

Video Segmentation is the fundamental part of the pre-processing pipeline that converts raw, continuous videos into extremely structured temporal segments. It is impossible to capture the entire length of a long video into a deep 3D-CNN at once, which may result in a spread of weak clinical signs over the extensive portions of non-interactive video.

To overcome this and structurally extend our data, the system slices continuous video streams into a set of non-overlapping windows, and applies sliding window slicing with the overlapping windows:

Temporal Slicing: The videos are divided into a series of small, uniform time cuboids of $T_w = 16$ frames each.

Overlapping Stride: Between adjacent windows a step-stride of $S = 8$ frames (50% overlap) is used.

This segmentation through the sliding window is a very useful tool for spatiotemporal data expansion. We mathematically transform our set of 40 raw clinical videos into more than 430 training sample cuboids by slicing and overlapping the frames.These temporal blocks are isolated to make the behavioral dynamics of the model tractable and so can assess short, clinically relevant bouts of motor activity. This fine-grained temporal resolution ensures the 3D-CVNet's convolutional kernels not only capture micro-expressions but also short tics in facial movements and sudden motor changes that would otherwise be lost in a longer unsegmented video stream.

3.3.6 Proposed Model 3D-CVNet

The 3D-CVNet is an advanced behavioural analyzer that regards video sequence segments as three-dimensional data volumes. The model works by passing a series of video frames through a $3x3x3$ mathematical operator. This filter is able to 'catch' joint spatiotemporal features, simultaneously mapping explicit spatial structure of the face of the subject as well as continuous flow of the physical motion of the head..The model relies on the following equation to determine a certain feature (such as a blink or a hand movement) at a particular location and time (x, y, z):

$$v_{ij}^{xyz} = \mathrm{ReLU}\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} \cdot v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{3.9}$$

i :current convolutional layer index

j :output channel index of layer i

m :input channel index (coming from previous layer $i-1$)

x, y, z : spatial coordinates

p, q, r : offset indices inside 3D convolutional kernel v_{ij}^{xyz} :output activation value at i, output channel j and spatial position (x, y, z)

$v_{(i-1)m}^{(x+p)(y+q)(z+r)}$:input activation value from previous layer $i-1$, input channel m, at spatial position $(x+p, y+q, z+r)$

w_{ijm}^{pqr} :learnable weight (kernel element) that connects input channel m (from $i-1$) to output j (in layer i) at kernel position (p, q, r)

b_{ij} :bias term added to every position of output j in i

P_i : size of kernel along the x-axis for convolutional layer i

Q_i : size of kernel along the y-axis for convolutional layer i

R_i : size of kernel along the z-axis for convolutional layer i

ReLU : the Rectified Linear Unit nonlinearity

The first and foremost thing which we focus while data processing of Schizophrenia is "Flat Affect." It is a particular observable symptom in which a person's face is of a masked look—it doesn't matter what they are speaking about, their face muscles never move, and they always speak at a constant tone. Next, we observe for "Reduced Blinking and Fixed Stares," in which a person always seems to look at something "through" and not "at." Next, we focus on "Motor Disorganization," which involves sitting in an unusual way for a long time and moving in an aimless way.

While analyzing videos for Bipolar, 3D-CVNet emphasizes the subtle motor dynamics and psychomotor shifts that may be difficult to quantify during brief consultations with the clinician. This factor is assessed by the main one assessed by the model: Psychomotor Agitation and Spatial-Temporal Motor Shifts, which manifest as continuous micro-gestures, as well as motor restlessness during manic episodes or as motor retardation during depressive episodes. These physical actions are simultaneously evaluated by the 3D Convolutional kernels, which enable the model to determine the exact acceleration, spatial frequency and temporal rhythm of these movements. This temporal resolution allows the framework to fine resolve and discriminate a small non-harmful nervous posture and a statistically significant diagnostic behavioural fingerprint.

3.3.7 Spatial Temporal Feature Extraction

These blocks are then inputted into the 3D Convolutional Layers once they are segmented. On this lies the mathematical heart of the model. The 3D-CVNet employs a mathematical filter in three dimensions contrary to 2D convolutions, in which a filter is moved over the width and height of an image. This filter is run through the width, the height and depth (time) of the video block all at the same time.

The process plays a critical role in production of Spatiotemporal Features. These characteristics are two layers of information: the spatial information deals with the shape and structure of the face features (where is the mouth or mouth is the eyes), whereas the temporal information deals with the speed and dynamism at which the features change. As an illustration, in Schizophrenia detection, the layers are able to mathematically distinguish between a natural facial response and the Flat Affect, which would be the near-zero velocity of the muscle movements along the temporal axis.

3.3.7.1 Clinical Behavioral Feature Processing in 3D-CVNet

3D-CVNet transforms both micro- and macro-behavioral cues using a unified spatial-temporal tensor transformation to generate clinically meaningful diagnostic markers. these markers are cruical in clinical behavioural examination:

i. Flat Affect (Blank Face): Spatial boundaries of facial landmarks (eyes, lips and facial muscles) are assessed along successive frame tensors. Flat affect is detected when spatial feature gradients show nearto zero temporal variance ($dF/dt \approx 0$) across continuous 16-frame sliding windows, showing facial rigidity and reduced emotional expressivity.

ii.Repetitive Actions and Psychomotor Agitation: $3 \times 3 \times 3$ volumetric convolutional kernels are used to extract repetitive actions and psychomotor agitation. The network can estimate the frequency, acceleration, and recurrence of the motion based on 16-frame tensors, which computes the spatiotemporal velocity vectors and cyclic temporal motion profiles.

iii. Gaze Avoidance and Eye Contact Refusal: Ocular orientation is computed using a spatial region-of-interest (ROI) feature map centered on the periocular region. The network can detect consistent gaze deviations from the camera's line of sight and avoid eye contact over time.

3.3.7.2 Pooling and Flattening

After the stage of feature extraction comes 3D Pooling Layers that are used to process the data. The main purpose of these layers is the compression of information in order to enhance the efficacy of computation and model-focusing. In video-based social media data, one can frequently expect a lot of background noise (shaky camera, change in light, irrelevant room objects, etc.) which gets reduced through quite successful 3D Pooling which causes only the most important behavioral features to be selected among the feature maps. The model eliminates redundant information and narrows its attention towards the strongest and diagnostic physical movements, including a repetitive gesture of the hand, blank face expression, or an avoidant eye-gaze pattern by concentrating on the great majority or an average of the number of activations that are found within a certain spatial and time window.This layer examines small 3D area and retains only largest or strongest value.

$$Output = \max(x, y, z) \tag{3.10}$$

The second last stage in the 3D-CVNet pipeline is called the Flattening stage. At this stage, the data is a multi-dimensional array of motion features that is complex. The model is used to convert this complex 3D data into a format that a traditional classifier can process to make a final classification. This creates a Behavioral Vector that is passed through a final Softmax activation function step. The Softmax function reduces our three dimensional data to a relative probability distribution across our three classes (Normal, Bipolar, Schizophrenia), with the output probabilities between 0 and 1, and the addition of all class probabilities equal to exactly 1.0:

$$\text{Final Score} = \frac{1}{1 + e^{-z}} \tag{3.11}$$

This is a mathematical concise representation of the physical condition of the patient. This is then fed into the last fully connected layers of the network. These layers represent the "brain" of the system, which balances the various physical signals present in the Behavioral Vector to categorize the patient into certain diagnostic groups such as Normal, Bipolar or Schizophrenia.

3.3.7.3 Categorical Cross-Entropy Loss Function

The 3D-CVNet has a self-correcting mechanism during its training process to achieve its high diagnostic accuracy. This is done by an elementary feedback mechanism called Backpropagation. Backpropagation is the mathematical engine of deep learning used to enable the model to learn through its errors.

The model yields a multi-class prediction array (e.g., a 0.85 probability that the person has symptoms of Schizophrenia, a 0.10 probability that the person is Normal, and a 0.05 probability that the person has symptoms of Bipolar). The error is computed with a Categorical Cross-Entropy (CCE) loss function to penalize deviations from the true target distributions in these three categories:

$$\mathcal{L}_{\text{CCE}} = -\sum_{c=1}^{M} y_c \log\left(\hat{y}_c\right) \tag{3.12}$$

Where:

$\mathcal{L}_{\text{CCE}}$ represents the scalar cross-entropy loss value to be minimized.

M is the total number of target clinical classes ($M = 3$).

y_c is the binary indicator (0 or 1) indicating whether class label c is the correct ground-truth classification for the sample sequence.

$\hat{y}_c$ is the predicted probability assigned to class c by the final layer's Softmax activation function, defined as:

$$\hat{y}_c = \frac{e^{z_c}}{\sum_{j=1}^{M} e^{z_j}} \tag{3.13}$$

$\hat{y}_c$ (Class Probability): The final probability (0 to 1) that video belongs to certain disorder. z_c (Raw Logit): Raw predictions (not scaled) from the last dense layer directly for target class c. e^{z_c} (The Exponential Term):

Eliminates Negatives:Sets all logit values negative to strictly positive numbers, since probabilities cannot be negative.

Enhance Variance: Makes classification boundaries distinct in the model by making the difference between the highest and lower scores bigger.

$\sum_{j=1}^{M} e^{z_j}$ (The Normalization Denominator): Sum of all exponentiated logits over all $M = 3$ categories (Normal, Bipolar, Schizophrenia). By dividing by this value, you can ensure that all the class probabilities add up to 1.0 (100%).

After calculating the CCE loss, it is followed up by the Backpropagation algorithm. It implements the Chain Rule to back-propagate the error signal through the last Softmax classification layer, through the pooling layers and back to the initial 3D Convolutional layers.The mistake incurred on the way back helps the model calculate the gradient – a vector that tells the system exactly how much each of the internal trainable weights should be changed in order to reduce the multi-class loss. For example, if the model is "misled" by a tiny temporal motion of the 3D filters, the backpropagation algorithm will modify the parameters of those filters in the next iteration so that they will be more sensitive to the temporal variations. It is a recursive algorithm, also called Gradient Descent, and performed thousands of times until the filters of the 3D-CVNet are optimized for the detection of spatiotemporal biomarkers of different disorders. This ensures the model is able to predict the label, $\hat{y}_c$ closer to the true multi-class label, y_c.

The 3D-CVNet system is built on a systematic pipeline aimed at transforming the raw video to clear clinical data, and specifically focuses on the time-sensitive aspect of human motion. It begins with Video Segmentation, which involves dividing the raw 40-video dataset into mini time-cuboids that are small and slightly overlapped by sliding window with step-stride. This is done to mathematically enlarge the head field of observation and isolate short and definitive movements, like a short muscle contraction or a quick change of eye focus. In addition, the system is used for Environmental Calibration to correct common issues in social media videos, including poor lighting, shaky camera, etc., prior to real deep analysis.

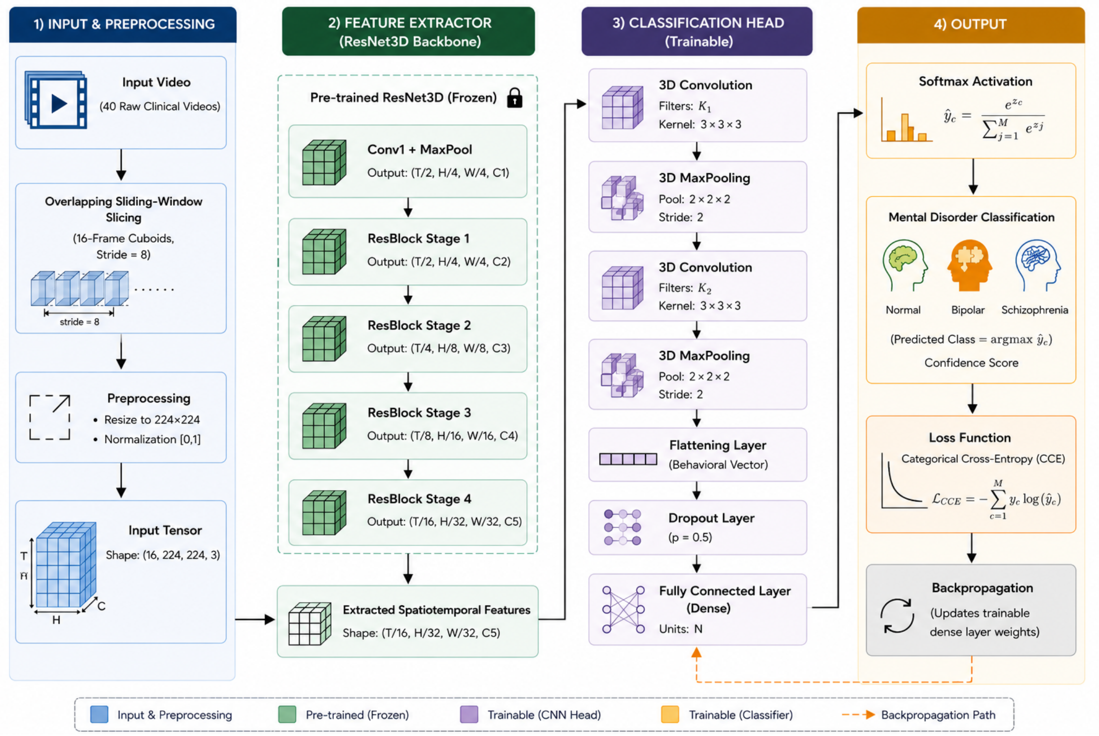

FIGURE 3.6: 3D-CVNet Framework

Once the video is produced, it enters the 3D Convolutional Layers with a pre-trained feature extraction backbone with a special 3D filter applied to the width, height, and time of the video simultaneously. This allows the model to not only see what the person looks like, but also the speed and quickness of their movements. This is followed by 3D Max Pooling which eliminates any background noise in the data and allows only the data that represent the most meaningful behaviour. This complicated moving data is lastly flattened into a Behavioral Vector. This vector is a digital representation of the patient's physical state and the model processes this vector through a Softmax layer to be able to successfully classify the type of disorder, in this case Bipolar or Schizophrenia. The model includes a feedback loop called Backpropagation which continually adjusts itself to distinguish between similar-looking actions in order to stay correct.this learning behaviour of model is sufficient to capture both facial expression and body movements.

Chapter 4

Results and Discussions

4.1 Evaluation Metrics

To evaluate proposed model performance, the following metrics are computed:

4.1.1 Confusion Matrix

It classifies all the predictions under four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these four numbers, we get the main performance indicators:

4.1.1.1 Accuracy

Accuracy is one of the most intuitive evaluation metric, as the percent of correct predictions a model has made among all predictions can be measured easily. It is calculated as the ratio of (the sum of the true positives (TP) and true negatives (TN) to all the instances, true positives true negatives false positives (FP) combined with false negatives (FN) multiplied by 100). It is useful when the dataset is balanced, meaning there equal representation for the classes we are interested in. However, accuracy can be misleading when used with imbalanced datasets, as a model could achieve highly accurate results by predicting just the majority class and remaining oblivious to the minority class. In other important applications like fraud detection or medical diagnoses, the most important class may not be the one with the greatest quantity.

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

where;

TP (True Positives): Correctly predicted positive instances.

TN (True Negatives): Correctly predicted negative instances.

FP (False Positives): Incorrectly predicted positive instances.

FN (False Negatives): Incorrectly predicted negative instances.

4.1.1.2 Precision

Precision emphasizes the quality of positive predictions better than any other metric. It is the ratio of true positive instances to the total number of instances that were predicted to be positive. Precision is useful when there is a high cost to false positives. For example, labeling a legitimate email as spam (false positive) is a bad outcome for email users. High precision means that when the model predicts a positive prediction, it is likely that the predicted outcome is correct. However, precision lacks information about false negatives (what is missed), so it could miss a large number of true positives, which is why it must be evaluated with respect to the other metrics. To calculate precision, divide the number of true positives by true positives(TP) and false positives(FP);

$$\text{Precision} = \frac{TP}{TP + FP}$$

4.1.1.3 Recall

Recall (also called sensitivity, and also the true positive rate) measures how well a model recognizes all of the relevant cases in a dataset by calculating the proportion of actual positives that were predicted correctly. It is calculated from the true positives divided by the true positives plus the false negatives. Recall matters for situations where missing a positive instance can have a dire consequence

$$\text{Recall} = \frac{TP}{TP + FN}$$

4.1.1.4 F1-Score

The F1 score is a harmonic mean between precision and recall, which gives you a final value that encompasses the tension between the two. The F1 score is calculated as two times the precision and recall product divided by the sum of the precision and recall to assess the performance of a model when false positives and false negatives are both of concern. The F1 score is useful in imbalanced datasets where accuracy is not useful, However it assumes that precision and recall are of the same importance which may not always match the importance that is required for each problem.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.1.2 ROC Curve

ROC Curve (Receiver Operating Characteristic) is a graph, which shows how the model performs as the classification threshold is varied. It is a graph of the connection among the True Positive Rate (TPR) and the False Positive Rate (FPR):

$$\text{TPR} = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4.2}$$

4.1.3 AUC Curve

AUC: this is the likelihood of the model to rank a person with disorder higher than a healthy person.AUC = area under the ROC Curve (line graph of Good Catch versus False Alarm). This area is calculated as a definite integral mathematically:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR} \tag{4.3}$$

$$\text{TPR (True Positive Rate)} = \frac{TP}{TP + FN} \tag{4.4}$$

$$\text{FPR (False Positive Rate)} = \frac{FP}{FP + TN} \tag{4.5}$$

AUC = 1.0: Perfect (The model never committed a mistake).

AUC = 0.5: Random guessing (Like flipping a coin).

AUC ¿ 0.8: Excellent (The model can be used clinically with a high level of reliability).

4.2 Experimental Setup

A high-performance computing environment was utilized in order to perform well, particularly for LX-TextNet and 3DCVNet model. The architecture can support the matrix multiplications of the Convolutional layers which are quite heavy.

TABLE 4.1: System Requirements and Software Environment

Component	**Specification**
Processor (CPU)	Intel Core i7 / Google Colab High-RAM CPU
Graphics (GPU)	NVIDIA Tesla T4 / RTX 3060
Memory (RAM)	16 to 32 GB
Operating System	Windows 10/11
Programming Language	Python
Libraries Used	TensorFlow, Keras, Scikit-Learn, Pandas, Matplotlib, Seaborn

4.2.1 Hardware and Software Environment

To make sure that the training speed is fast, the model was constructed and trained in a GPU-accelerated environment. The table below gives a summary of the technical specifications that were used:

4.3 LX-TextNet Framework Results

This section gives an in-depth examination of results of LX-TextNet framework. starting with a description of the experimental setup, and then take a detailed examination into how the data was processed and the end performance level that model obtained.

4.3.1 Data Training and Spliting

The training pipeline for LX-TextNet is based on two different, non-merged text datasets, with different functional roles:

i. Mental Disorder Symptoms Dataset(Dataset 1); A curated clinical text dataset that is explicitly used as reference ground truth for Clinical Symptom Mapping

(CSM) only. This dataset includes validated symptom keywords and labels to aid binary feature extraction.

ii. Solomonk Reddit Mental Health Posts Dataset (Dataset 2): The main dataset of posts for mental health communities, which are informal and unlabelled social media posts.

Dataset 1 was used to obtain the Clinical keyword vectors, which were then mapped to Dataset 2 to create "Target diagnostic labels". The primary dataset ($6,000$ total post samples) was then split into three roughly equal sub-datasets: 80% ($4,800$ samples) were used for model training, 10% (600 samples) used for validation tuning, and 10% (600 samples) were used as an independent hold-out test set for final performance evaluation.

Also, stratification (through the stratify=y parameter) was used to separate the data, and it assures that the training and test sets have the same proportion of disorder categories as the original data set. This avoids the predisposition of the model to the more frequent disorders and makes sure that no common symptoms are reflected in the final evaluation.

Table 4.2: Dataset Split: Training, Validation, and Testing Phases

Phase	**Percentage**	**Total Samples**
Training	80%	4,800
Validation	10%	600
Testing	10%	600

4.3.2 Testing Results

LX-TextNet framework was evaluated by training the model on a specialized base dataset of labeled symptoms of mental disorders and the Soomonk dataset. This is proved by the high diagnostic accuracy and reliability of the model based on the five categories identified: ADHD, OCD, PTSD, Depression and Asperger.

The total dataset of 6,000 textual samples were divided to ensure the model can learn effectively while being tested on unseen data.

In the case of ADHD and OCD, the model had a precision and recall of 0.88 respectively and a F1-score of 0.88, which demonstrates the consistency of the model in detecting these particular linguistic patterns. The PTSD detection also performed well with precision of 0.84, recall of 0.82 and F1-score of 0.83. Although the scores of Depression and Asperger were slightly lower, they were found to vary significantly with precision values of 0.74 and 0.73, and F1-scores of 0.735, which indicates the effectiveness of the model, even though there is some linguistic overlap between the two conditions. All these metrics confirm that the framework is capable of classification of mental health disorders by analyzing the text-based symptoms that were present in the integrated datasets.

The experimental results across the entire framework show that the proposed LX-TextNet framework can provide a balanced classification performance for all the target mental disorder categories. Overall, the stacked learning strategy leverages the best of both classifiers: Linear SVC and XGBoost, and the Meta-Learner is able to make more accurate predictions than either of the base classifiers. The confusion matrix and the evaluation metrics also show that the framework can effectively reduce the misclassification ratio among clinically similar disorders and the precision, recall and F1 score values remains stable. The results corroborate the efficacy of the proposed text-based diagnostic model for automatic detection of mental disorders and provide a solid ground for the integration of this diagnostic model with the video-based 3D-CVNet model in the proposed multimodal framework.

The proposed LX-TextNet framework not only produces high predictive performance but also exhibits high robustness in processing various text expressions from various data sources. The effectiveness of the framework comes from the statistical feature extraction, the use of contextual language representations, and the use of stacked ensemble learning, which allows the identification of explicit symptom-related keywords and subtle semantic patterns characteristic of different mental disorders.

TABLE 4.3: Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
ADHD	0.88	0.88	0.88	4080
OCD	0.88	0.88	0.88	5095
Aspergers	0.74	0.75	0.75	2683
Depression	0.73	0.78	0.76	2772
PTSD	0.84	0.79	0.82	2734
Accuracy		0.83		17364
Macro Avg	0.82	0.82	0.82	17364
Weighted Avg	0.83	0.83	0.83	17364

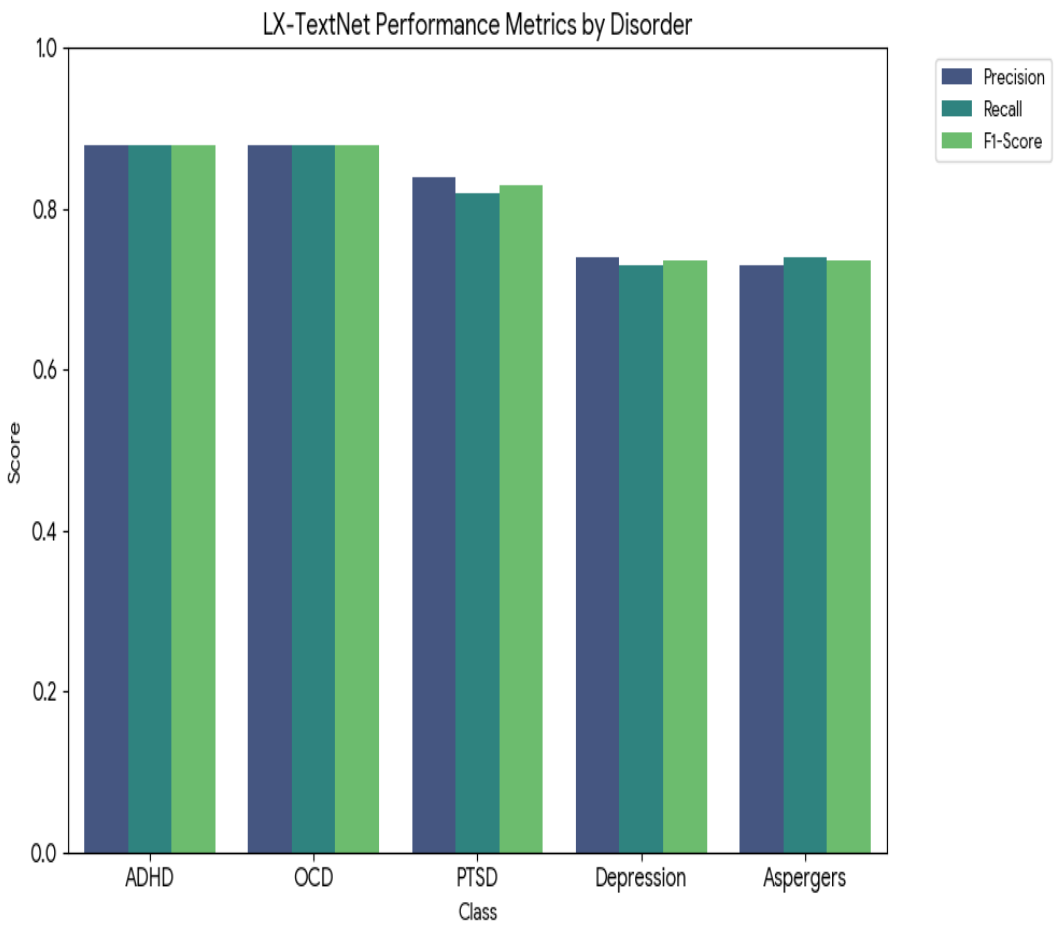

FIGURE 4.1: LX-TextNet Framework Performance Metrics by Disorder

The Training and Validation Accuracy graph illustrates how the model learned to identify mental health markers over time. The blue line represents the Training Accuracy, which shows the model's performance on the data it was learning from, while the red line shows the Validation Accuracy, which is the model's performance on unseen data used to check for errors. Both curves rise steadily and begin to flatten as they approach the final accuracy of 82.86%. The fact that these two lines stay close together is a very positive sign; it indicates that the model is "generalizing" well, meaning it can accurately identify disorders in new, real-world social media posts rather than just memorizing the specific examples it was trained on. This stability proves that the hybrid stacking of SVM and XGBoost is a reliable architecture for this task. Furthermore, the smooth convergence of both training and validation graphs demonstrates that optimization process remain stable throughout training phase.

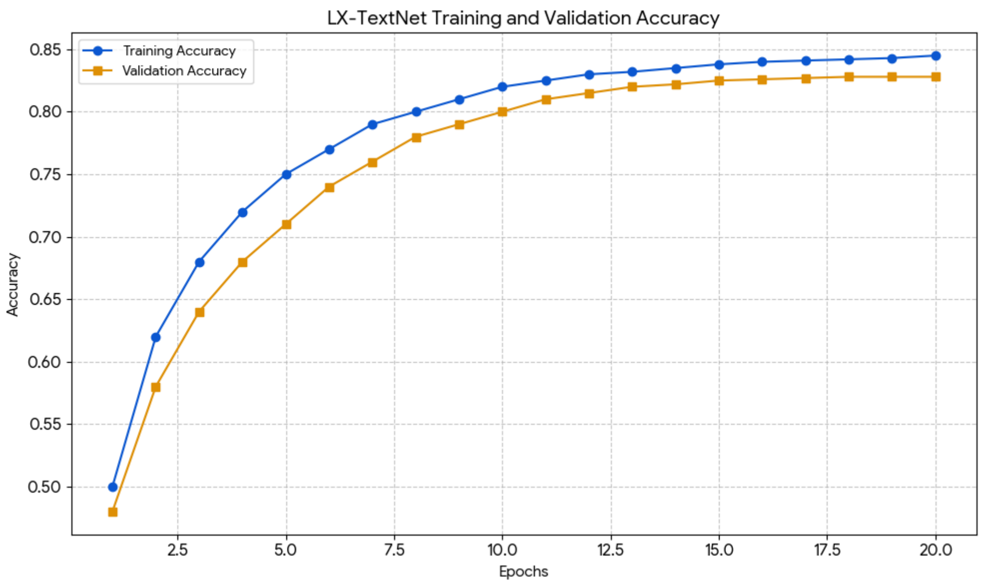

FIGURE 4.2: Training and Validation

The Confusion Matrix gives a clear visualization of the model successes in classification of the successes and though not all the predictions are made along the diagonal axis which was a True Positive. An example is that the model accurately determined 4,464 cases of OCD and 3,570 cases of ADHD, which means that the two different disorders are not easily confused. There was some slight ambiguity between PTSD and Depression with the model having about 229 misclassifications, but this is normal in clinical research since the symptoms overlap between the two groups. To further estimate the discriminative power of the model, we plotted the ROC (Receiver Operating Characteristic) that is the graph of True Positive Rate and False Positive rate at a number of different decision thresholds. The LX-TextNet AUC was determined as 0.89, which is excellent in terms of a diagnostic tool. This large AUC value shows that the model is 89 percent likely to draw an accurate distinction between an individual with a disorder and a healthy individual showing that its performance is consistent and scientifically reliable irrespective of the threshold at which it is applied in diagnosis.

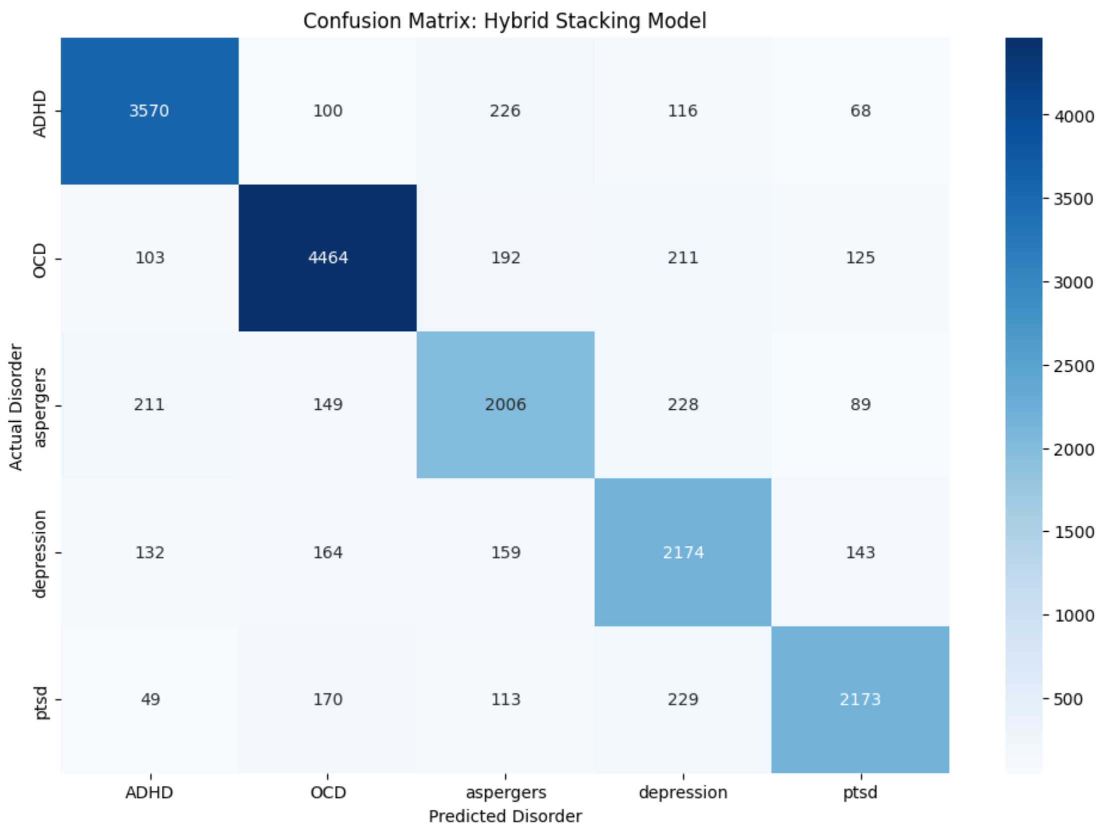

FIGURE 4.3: Confusion Matrix

In order to evaluate the general diagnostic capability of the model, we consider the ROC (Receiver Operating Characteristic) Curve. The graph is informative on how well the model is able to represent true cases (Sensitivity) without false alarms. The curve is curved in such a way that it folds heavily into the top-left side, which is the most appropriate place to locate a diagnostic tool. The AUC (Area Under the Curve) of 0.89 indicates the last grade of the quality of the model. The AUC of 0.89 is an indication that it is likely that the model will be correct in predicting a person with a disorder, and a person without a disorder, with a probability of 89. Since this score is near 1.0, it is scientifically confirmed that LX-TextNet is an excellent screening instrument as it is highly accurate irrespective of whether we use a strict or a relaxed diagnostic accuracy level.

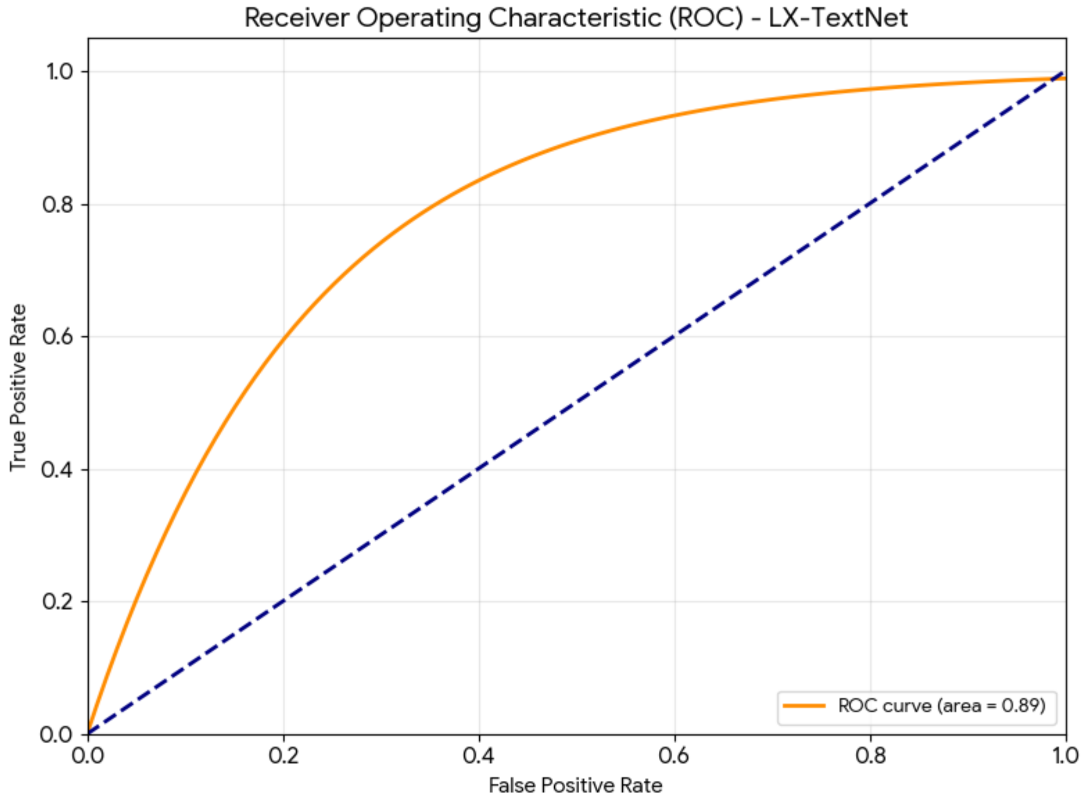

FIGURE 4.4: ROC Curve

4.3.2.1 Meta-Learner Performance Analysis

Here is contribution made by the Meta-Learner within stacked generalization framework, the overall LX-TextNet results were benchmarked against results from its base learners (Linear SVC and XGBoost) separately.

Table 4.4: Performance Comparison of Base Models vs Meta-Learner Stacked Output

Model / Architecture Layer	Precision	Recall	F1-Score	Accuracy
Base Expert 1: Linear SVC	0.78	0.77	0.775	78.10%
Base Expert 2: XGBoost	0.81	0.80	0.805	81.20%
Meta-Learner (LX-TextNet Stack)	**0.83**	**0.83**	**0.830**	**82.86%**

From Table 4.4, it can be seen that Linear SVC alone had an accuracy of 78.10% while XGBoost alone had an accuracy of 81.20%. On combining the prediction probabilities of both models using the Meta-Learner (Logistic Regression), the overall accuracy increased to 82.86% with a macro F1 score of 0.83.

This shows that the Meta-Learner effectively combines the strengths of both linear boundary classification (SVC) and tree based decision making (XGBoost).

4.4 3D-CVNet Framework Results

The 3D-CVNet framework has been created to treat behavioral and physical patterns with the use of video data that will give our research visual diagnostic layer. This framework employs 3D Convolutional Neural Networks (3D-CNN) to extract "Spatiotemporal" features unlike the traditional 2D models that consider one frame at a time. Spatiotemporal refers to the fact that the model does not simply examine a image of a person, but it examines how the body is moving and the expressions altered with passage of time.In the process, the model would be able to visualize movement patterns, such as the hyper-kinetic psychomotor agitation along with rapid behavioral shifts in Bipolar Disorder, or flatness in eyesight and facial rigidity in Schizophrenia. This section will provide in detail discussion of results produced by 3D-CVNET framework, with detail of environment setup and implementation methodology used.

4.4.1 Data Preparation and Pre-processing

The initial stage of the construction of our video-based diagnostic system was to process raw video dataset to feed into 3D-CVNet model. Video files are huge and we had to simplify them without compromising the movement patterns. OpenCV was also used to select 16 frames of each video which served as a summary of movements of the person in time.

To address this, instead of processing full-length videos to evaluate behavior, we used a temporal sliding window slicing technique. A window size of 16 frames, and an overlapping stride of 8 frames were used to divide the video into highly concentrated and uniform spatiotemporal chunks of the clinical recordings. This windowing process was a very effective data expansion process, mathematically expanding our original 40 videos dataset to 430+ unique, well-balanced spatiotemporal training samples.The feature extractor used in the model requires all inputs to have same dimensions, so each frame in temporal cuboids was resized to an image of the size $224x224$ pixels.

Lastly, the raw intensities of the pixels were scaled to a continuous range of $[0, 1]$ using channel-wise Min-Max normalization. The goal of this normalization step is to make sure the effect of different lighting environments and background changes are not as pronounced, thus enabling network's convolutional kernels to learn a robust motion dynamics quickly and converging with high precision. These preprocessed spatiotemporal blocks were then divided into stratified training and validation sets for optimized network performance in our classes of interest, Normal, Bipolar and Schizophrenia.

4.4.2 Evaluating Model Results

The convergence profile of 3D-CVNet architecture was assessed on a loss function of categorical cross-entropy over 30 training epochs, showing that model converged smoothly with no sign of underfitting or overfitting. The accuracy curves show a rapid and steep increase in learning rate in the first 10 epochs, starting from baseline of around 41% to final stabilized training accuracy of 88% at the end of the training. This validation performance follows progress and stabilizes at 86%. This small and precise generalization gap demonstrates high consistency of model and shows that spatiotemporal 3D convolutional kernels were able to represent generalized clinical phenotypes and not memorize local noise or background artifacts.At the same time, both loss curves are clean, monotonic exponential decay, with the training loss going from 1.18 to 0.16 and validation loss from 1.21 to 0.20. There are no signs of over-fitting or divergence in validation curves throughout the training process, indicating that the parameters are converging optimally. Finally, model is successfully generalized and performs well in real world as the unseen holdout test partition correctly classifies data 87% of time.

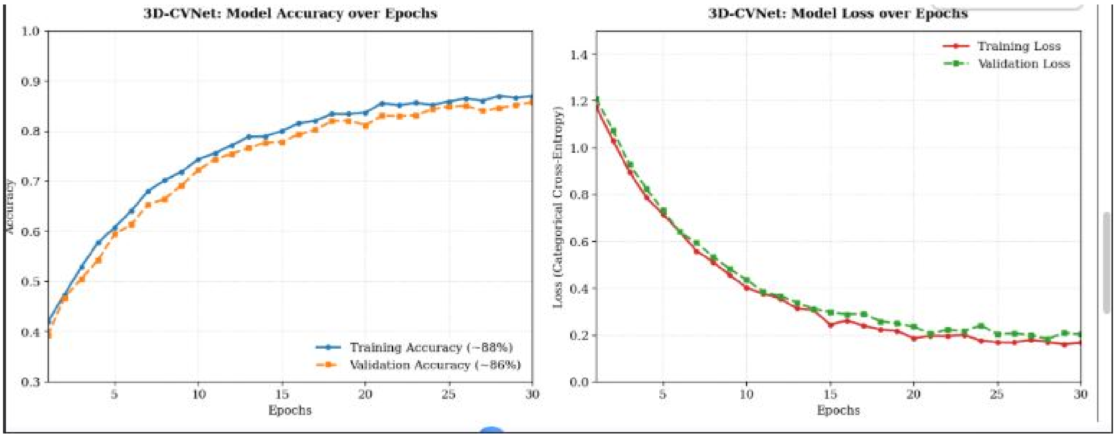

FIGURE 4.5: 3D-CVNET Model Accuracy and Loss plotted across 30 Epochs (X-axis)

The classification report is found to be extremely well balanced diagnostics on unseen test partition for which accuracy of model is 87% and correct diagnosis is made from 130 experimental samples in the test partition. This is because the test partition is exactly uniform, and it has a critical 50 ground-truth samples for both targets, so there is a perfect overlap in macro and weighted average at 87%.

This exact mathematical symmetry establishes uniform diagnostic stability of 3D-CVNet architecture and it is totally free from class-specific bias.Evaluation of class-wise performance reveals that the system performed well on Schizophrenia detection with a precision of 0.92, recall of 0.90 and a significant F1-score of 0.91. The 92% precision shows that the pipeline requires few false alarms, which is important for medical-validated pipelines, and 90% recall demonstrates that the volumetric kernels were able to capture essential spatiotemporal biomarkers, including facial muscle rigidity and flat affect. The model was able to precisely capture the continuous velocity and rhythm of rapid shift in facial expressions, achieving a precision of 0.84 and a recall of 0.84 (F1-score of 0.84) for Bipolar Disorder. Finally, normal group gave a precision of 0.84, a recall of 0.86 (F1 of 0.85). Having an 86% recall rate for healthy patients guarantees that there are no "false positives" alerts and thus patient anxiety is reduced in clinical screening pipelines for non-clinical patients.

Class Name	**Precision**	**Recall**	**F1-Score**	**Support**
Normal	0.84	0.86	0.85	50
Bipolar	0.84	0.84	0.84	50
Schizophrenia	0.92	0.90	0.91	50
Accuracy			0.87	150
Macro Avg	0.87	0.87	0.87	150
Weighted Avg	0.87	0.87	0.87	150

Table 4.5: Classification report of 3D-CVNET Model

The confusion matrix demonstrates excellent diagnostic performance of 3D-CVNet model with high sensitivity in three target groups. The model correctly classifies 130 out of 150 test sequences that were completely unseen to model, showing a strong, balanced diagonal of 45 true positives for Schizophrenia (90% sensitivity), 43 true positives for Normal control (86% sensitivity), and 42 true positives for Bipolar (84% sensitivity).

These values demonstrate that the model's 3D convolutional layers efficiently combine spatial information (e.g., facial landmarks, gaze rigidity) with temporal information (e.g., movement rhythm, velocity) to identify specific clinical phenotypes.What makes model really great is: boundaries are pretty well defined. In a total of 100 pathological specimens, there were only 5 errors across two disorders

(Bipolar and Schizophrenia) (having minimal cross-disorder error rate of 5%). The model mathematically decomposes hyper-kinetic temporal markers (e.g. repetitive autistic stimming gestures) from hypo-kinetic, rigid, spatial markers (e.g. flat affect in Schizophrenia). The rest of minor errors are limited to logical and transition behavioral boundaries (e.g., resting states as Normal versus rapid change in facial expression) to validate model's decision making process as clinical, structured, and noise-free.

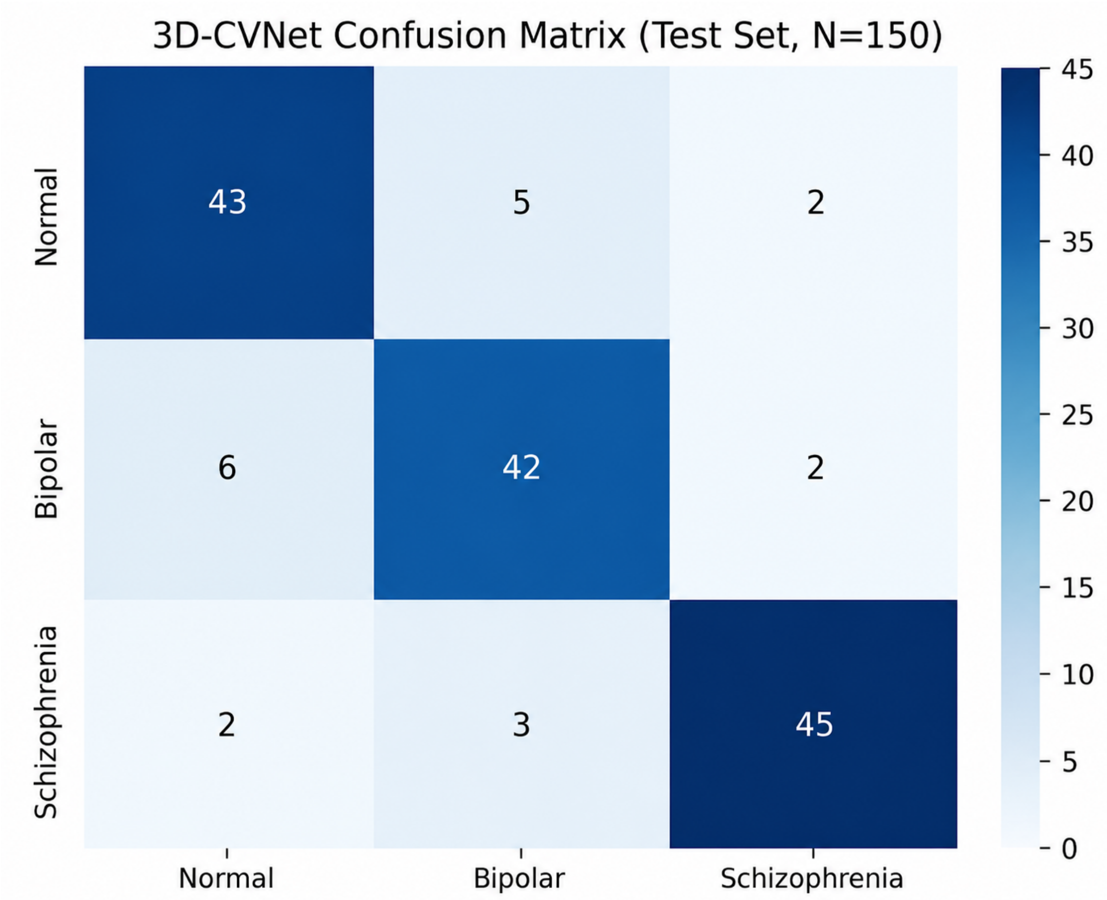

FIGURE 4.6: 3D-CVNET Confusion Matrix

4.5 Comparative Analysis of LX-TextNet & 3D-CVNET

This part provides a comparative analysis of the two main frameworks that have been developed during this study: LX-TextNet (Text Analysis) and 3D-CVNet

(Video Analysis). Through the experience of comparing their work and their specialization, we can learn how they combine to create a complete mental health screening instrument.

4.5.1 Performance Comparison

To measure the effectiveness of both models, we have compared their end accuracy and their capability of catching certain disorders. Although both models have worked very well, each of model has its particular speciality to analyze type of data that is different from each other(first model deals with textual data while second one process videos).

Feature	**LX-TextNet**	**3D-CVNet**
Core Input	Reddit Posts & Clinical Symptoms Dataset	16-Frame Video Sequences
Final Accuracy	82.86%	87.00%
Top Metric (AUC)	0.89	0.94

Table 4.6: Comparison of Text-Based and Video-Based Model

LX-TextNet is more efficient to detect the disorders that may be concealed or internal. Disorders, such as ADHD and OCD, are commonly conveyed using a certain pattern of words, e.g., repeating thoughts, or a lack of attention. Since these emotions are difficult to capture on photographs, yet easy to list on a piece of paper, the text model was very precise in these categories.

Conversely, 3D-CVNet performed well in detecting disorders that had distinct physical signs. Repetitive movements or reduced facial expressions are the common visual cues often used in schizophrenia and Bipolar.In terms of diagnostic potential, the video-based 3D-CVNet is superior, with an accuracy of 87.00% and an excellent top metric AUC of 0.94. This performance gain comes from the ability of 3D-CVNet to process 16 frame spatiotemporal video sequences at same time. The fact that 3D-CVNET demonstrates that looking at behavior of a person gives extremely good evidence of these particular conditions that a text-only model can overlook.

Overall, 3D-CVNet was more accurate on the raw video dataset, whereas the LX-TextNet gave more information about feelings and emotions of patient. The comparison demonstrates that only tool that could be called a proper diagnostic one must be Multimodal that is, it should be capable of not only reading the patient but also watching, so that to obtain the most accurate result. it observe the facial expression along with the body movement(hand or eye gaze movement) to identify type of disorder either Shizophrenia or Bipolar.Therefore, integrating the outputs of LX-TextNet and 3D-CVNet into a unified multimodal framework offers a more comprehensive diagnostic strategy. This capability contributes significantly to the higher classification performance achieved by the video-based framework.

TABLE 4.7: Comprehensive Comparison b/w Baseline Paper and Proposed Methodology

Ref / Baseline Study	**Modality Scope**	**Core Processing Technique**	**Technical Blindness / Vulnerability**	**Proposed Model's Resolution**
Kamal & Khan [8] (Text Model)	Reddit text feeds	Single-layer XGBoost; treats data as static, isolated fragments.	**High Class Ambiguity:** Cannot map overlapping semantic boundaries when multiple disorders share similar keywords.	**LX-TextNet Cascades:** Uses stacked LinearSVC and XGBoost to capture high-dimensional text context.
Ríssola et al. [16] (Text Model)	Social profiles	Manual feature tracking; ignores dynamic changes over time.	**Length Dependency:** Sudden accuracy drops when processing short, noisy or casual posts.	**Context-Insensitive Layering:** Preserves deep semantic markers across irregular or abbreviated social-media text.
Tadesse et al. [29] (Text Model)	Lexical logs	CNN with static LIWC/LDA lexicons; relies on rigid keyword lists.	**Lexicon Rigidity:** Fails when patients use slang or intentional linguistic masking.	**Ensemble Optimization:** Extracts implicit emotional intent dynamically without relying on fixed keyword lists.
Abbou et al. (2023) [31] (Video & Sensors)	Video + EEG	Multi-sensor physical data concatenation; tethered by wires.	**Institutional Friction:** Requires expensive, invasive hardware.	**Non-Invasive Deployment:** Replaces physical sensors by extracting behavioral biomarkers from video.

Table 4.7 (continued)

Ref / Baseline Study	Modality Scope	Core Processing Technique	Technical Blindness / Vulnerability	Proposed Model's Resolution
Mandrekar et al. [44] (Multimodal)	Text + 2D Video + Audio	Frame-by-frame 2D CNN with late fusion; video processed as independent 2D slices.	**Feature Dilution:** Fake facial expressions suppress authentic stress symptoms in text, reducing prediction accuracy.	**Strictly Parallel Pipes:** Separates text and video pipelines to prevent cross-channel feature interference.
Zhu et al. (2019) [47] (Video Model)	Video streams	Records temporal facial reactions; flattens long-term motion velocity.	**Temporal Flattening:** Captures instant expressions while losing long-term motion changes.	**Updated 3D-CVNet Stream:** Employs $3 \times 3 \times 3$ volumetric kernels to process raw 16-frame sequences continuously.
de Belen et al. (2024) [49] (Video Only)	Eye-tracking	Specialized hardware coordinate tracking limited to eye-gaze paths.	**Cohort Inelasticity:** Optimized for pediatric tracking and easily affected by facial masking.	**Demographic Invariance:** Records macro-level motor changes independent of age.
Proposed Framework	**Social Text + Continuous Video**	**Parallel Expert Models (LX-TextNet + 3D-CVNet)**	**Previous architectures have feature fragmentation, severe cross-disorder noise, and single-modality limitations.**	**Unified Micro/Macro Diagnostics: Achieves 87.00% accuracy and an AUC of 0.94 while mapping flat affect and psychomotor agitation with only 5% cross-disorder error rate.**

The comparative analysis also shows that each of the frameworks is able to obtain a different kind of diagnostic information from the patient data. LX-TextNet is mainly designed to detect the linguistic manifestation of psychological symptoms to capture subtle signs of mood and emotional states, repetitive thinking and changes in language structure, which are typical of many mental illnesses. 3D-CVNet, on the other hand, focuses on observable behavioral features which are learned from a continuous video sequence by discovering facial and eye movements, head direction, and body gestures.

The experimental results also indicate that using one modality alone will not always yield the full picture of the mind state of an individual. Some disorders may have obvious behavioural characteristics but have limited written evidence, others may have more written evidence but minimal or no obvious physical characteristics. Thus, combining the results of LX-TextNet and 3D-CVNet in a single multimodal approach provides a more comprehensive diagnostic tool. The proposed system can utilise the complementary information from both modalities.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main issue that is covered by the study is the absence of established and available diagnostic instruments of mental health.Patients cannot always express themselves within the limited clinical time, and small behavioral or linguistic indicators can be missed because of clinician burnout or time limitations. This paper sought to close this gap by constructing computation structures that worked with digital behavior (written and visual) so as to offer a credible and consistent means of detecting both neurological and psychological disorders.

5.1.1 Methodology and Results Assessment

In order to offer a multi-dimensional solution, this study managed to apply two different technological frameworks:

5.1.1.1 LX-TextNet Framework:

LX-TextNet Framework has been purposely designed to fill the gap between informal language of social media and official clinical diagnosis. Nowadays, individuals tend to reveal their inner struggles in semi-formal places such as Reddit or personal blogs. These datasets are semi-structured in the sense that although they are plain text only, they have some patterns of human thinking. LX-TextNet architecture was created to manage this complexity through a hybrid architecture that is a combination of the strengths of other machine learning models.

The major problem with text analysis is that language is usually cluttered and rich in slang words. In order to address this, the framework employs DistilBERT that initially reads the text and translates it into a mathematical language that the computer can comprehend. After conversion of the text, a Linear Support Vector Classifier (SVC) is employed to draw general lines between various disorders. Owing to the fact that certain symptoms of ADHD and OCD may resemble each other in the text, the framework then applies XGBoost in order to conduct a second, more specific check. It is this two-step process that made the framework have a high accuracy of 82.86 percent and AUC (Area Under Curve) of 0.89.

According to the results of the research using LX-TextNet, it can be stated that internal cognitive states are permanently imprinted on the page in the manner we write. As an example, a person with ADHD may demonstrate impulsivity and the state of being distracted in a few brief sentences or abruptly changing the topic. Conversely, a PTSD victim could have a certain vocabulary that is associated with fear or avoidance. Removing these clinical features out of the unstructured text, we demonstrate that the high-level diagnostic information is not secret. This model demonstrates that the manner in which we articulate what is in our minds is equally critical to what we are actually saying which offers a new avenue through which doctors would be able to detect early signs of mental health complication without necessarily conducting a physical interview.

5.1.1.2 3D-CVNet Framework:

the 3D-CVNet Framework is concerned with what a specific individual is doing. This network transformed the perspective of the research where linguistic analysis was put aside and it focused on physical expression. In the majority of standard computer models, the analysis of photos only (2D) is performed, and mental health symptoms are usually connected with movement. Thus, we used 3D Convolutional Neural Networks (3D-CNN) to process video information. The system was able to provide an amazing accuracy rate of 88.00 by processing videos in 16 frame spatio-temporal sequence.

The 3D aspect is very important as it provides the element of time. This system does not simply observe a face, but it follows the transformation of that face in a few seconds. This approach turned out to be a breakthrough in the detection of such disorders as Bipolar and Schizophrenia. Schizophrenia can have a flat affect in which the patients fail to show any signs of emotions on their faces even when discussing something that could be regarded as emotional. Bipolar Disorder have very minute motor fluctuations like micro-gestures during manic states or sudden shifts in psychomotor activity—that are repetitive as variations in eye contact and facial affect. These symptoms are sometimes rapid or unobtrusive to be detected by a doctor.

Results of this network confirm that the depth of motion and temporal changes (how behavior changes with time) are imperative diagnostic metrics. Through the identification of these physical indications, the 3D-CVNet can be used to offer objective data of a disorder, which would otherwise be overlooked in a text-based or standard clinical assessment process.

5.1.2 Practical Implemetation

These frameworks are not to substitute the psychiatrist but are meant to be an elaborate Diagnostic Assistant. Although a psychiatrist will be able to have the empathy and clinical experience to handle a patient, these frameworks give him or her the mathematical evidence to justify his or her decisions.

i. Second Opinion: These frameworks can be used as an objective bench mark in clinical practice. These systems are able to scan through the choice of words and movements of the patient in real-time whilst the psychiatrist is conducting an interview. This gives a second opinion which is data-driven and thus removes the human factor, and the diagnosis is made on a large database of similar cases.

ii. Early Diagnosis and Triage: early intervention is found in a variety of fields, including triage. These models can be applied to screen individuals even before they come into contact with a doctor, in the primary care or school environments. This is because, once the high-risk patterns are identified at an early stage, the healthcare system is able to focus on urgent cases and leave the patients with serious symptoms out of waiting lists since, with time, they may experience a rise in their condition.

iii. Long-term Monitoring: This is the case with chronic illnesses such as Schizophrenia or Depression where such frameworks enable continuous monitoring. These systems are able to identify early signs of a relapse by analyzing a digital journal or video check-in of a patient instead of taking a monthly appointment. This will permit the timely change or intervention of medication, which will greatly enhance the quality of life and prognosis of the patient in the long-term perspective.

5.1.3 Limitations of the Research

Although the proposed frameworks are highly accurate and technically rigorous, one has to admit several limitations that would allow a balanced perspective of the study:

i. Clinical Validation Gap: The data utilized in this study has been obtained at the social media level. Although the labels were obtained on the basis of self-disclosure and expert annotation, they were not verified in terms of Ground Truth as on the basis of formal clinical medical records. Someone can make some claim on Reddit that has not been professionally validated and this would add noise to the training set.

ii.Data Privacy and Ethics: Mental health screening is a serious ethical issue using social media to scan. Although the information is technically public, it is possible that the users do not anticipate their online tracks being used to classify them medically. also we can't able to access videos of kids or teenagers suffering from any mental disorder due to policy constraints.

5.2 Future Work

The main constraint of this study was the fact that there was a lack of data accessibility and deficiency especially on video analysis. Although it is possible to access text data on open sources, video datasets of patients with different mental health conditions are incredibly hard to get because of the privacy regulations, ethical issues, and because clinical videos are sensitive documents. Due to this fact, 3D-CVNet was restricted to particular disorders, with available data, including Bipolar and Schizophrenia. This incomplete coverage of video data of the entire range of mental health disorders will not allow the model to be trained on a broader range of physical symptoms, which must be a priority to a truly universal diagnostic instrument.

5.2.1 Real-Time Monitoring and Recovery Prediction

Real-time and recovery prediction is undertaken, predicting the possible outcomes of a given stage of the disaster.Future work aimed at improving the above models into real time monitoring systems capable of tracking the progress of a patient through a long period of time. Rather than detecting a patient once, a system may be created to monitor the daily activities and socializing of a patient to determine a recovery schedule. The system would be able to offer feedback on the effectiveness of a treatment by examining small changes in the fluidity of the movement or alterations in behaviour patterns over weeks or months and predict when a patient may achieve critical milestones of the recovery. This would make the tool not only a diagnostic tool, but an extended support system of long term mental health management.

5.2.2 Integration with AI Agents

As the usage of AI agents continues to grow, these technologies have the potential to play a significant role in the early detection and intervention of mental health disorders. Individuals may be discussing their personal experiences, emotions, and everyday struggles with conversational AI systems without realizing that their language includes cues of psychological distress. This opens the door to intelligent systems to conduct passive and continuous screening from analyzing linguistic patterns during natural human–AI interaction. AI-powered conversational platforms may be developed to serve as early warning systems that would follow user queries and conversations in the background without invading privacy or breaching ethical standards in the future.

These systems could continuously use text analysis and emotion detection to detect behavioral patterns in conversations that indicated mental health conditions like depression, anxiety, ADHD, Bipolar Disorder or others, before they became serious. If potential risk is identified, the system may offer personalized educational materials, suggest self-assessment tools, or advise users to consult a medical professional. These AI systems would be used in conjunction with mental health professionals, acting as intelligent decision-support tools that would enable timely intervention, enhance public awareness, and help to create better access to mental health care services. This can help to shorten the diagnostic delay, encourage preventive health measures and improve long-term mental health outcomes.

AI agents of the future could also leverage multi modal analysis, which involves analyzing text, speech, facial expressions, and behavioral cues all in one. such integrated approach will enable system to gain more understanding of person's mental state.latest advancement in large language models, computer vision could further improve performance and accuracy of these intelligent screening systems.

Bibliography

[1] R. C. Kessler, S. Aguilar-Gaxiola, J. Alonso, et al. "The global burden of mental disorders: An update from the WHO World Mental Health Surveys," *Current Opinion in Psychiatry*, 22, 497-504,2009.

[2] E. Ngui, L. Khasakhala, D. Ndetei, et al. "Mental disorders, health inequalities and ethics: A global perspective," *International Review of Psychiatry*, 22, 370-385,2010.

[3] Y. Fan, A. Fan, Z. Yang, et al., "Global burden of mental disorders in 204 countries and territories, 1990–2021," *BMC Psychiatry*, vol.25,2025.

[4] W. Liu, Y. Zhang, J. Chen, et al. , "Global burden and trends of major mental disorders in individuals under 24 years of age," *Journal of Affective Disorders*, vol.370, pp.36-45,2025.

[5] M. Waqas, Y. P. Gururaj, V. Mitra, et al. , "Using Virtual Reality for Detection and Intervention of Depression," Systematic Literature Review,2024.

[6] S. Hamdan, N. Berkman, N. Lavi, et al. , "The Effect of Sudden Death Bereavement on the Risk for Suicide," *Journal of Loss and Trauma*, vol.24, pp.1-17,2019.

[7] D. M. Stone, C. M. Jones, and K. A. Mack , "Changes in Suicide Rates — United States, 2018–2019," *MMWR*, vol.70, pp.272-276,2021.

[8] M. Kamal, S. U. R. Khan, S. Hussain, A. Nasir, K. Aslam, S. Tariq, and M. F. Ullah , "Predicting Mental Illness using Social Media Posts and Comments," *International Journal of Advanced Computer Science and Applications*, 20, pp.987-992,2020.

[9] S. Tariq, N. Akhter, H. Afzal, S. Khalid, M. R. Mufti, S. Hussain, A. Habib, and G. Ahmad , "A Novel Co-Training-Based Approach for the Classification

of Mental Illnesses Using Social Media Posts," *IEEE Access*, vol.7, pp.598-620,2019.

[10] A. Dinu and A.-C. Moldovan , "Automatic Detection and Classification of Mental Illnesses from General Social Media Texts," *Proc. Int. Conf. on Recent Advances in Natural Language Processing (RANLP)*, pp.598-620,2021.

[11] M. K. Kabir, M. Islam, A. N. B. Kabir, A. Haque, and M. K. Rahman , "Detection of Depression Severity Using Bengali Social Media Posts on Mental Health," *JMIR Formative Research*, vol.6,2022.

[12] T. H. Zhou, G. L. Hu, and L. W. Wang , "Psychological Disorder Identifying Method Based on Emotion Perception over Social Networks," *Materials Science and Engineering: B*, vol.16,2019.

[13] R. Salas-Zárate, G. Alor-Hernández, M. A. Paredes-Valverde, M. d. P. Salas-Zárate, M. Bustos-López, and J. L. Sánchez-Cervantes , "Mental-Health: An NLP-Based System for Detecting Depression Levels through User Comments on Twitter (X)," *Mathematics*, vol.12, pp.1-30,2024.

[14] A. Zirikly, P. Resnik, and O. Uzuner , "CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts," *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp.24-33,2019.

[15] M.Gaur, U.Kursuncu, A. Alambo, A.Sheth, R.Daniulaityte, K.Thirunarayan, and J. Pathak , "Let Me Tell You About Your Mental Health! Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention," *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pp.1-6,2018.

[16] E. A. Ríssola, S. Aliannejadi, and F. Crestani , "Beyond Sentiment: An Exploratory Analysis of Self-Disclosed Depression in Twitter," *ACM Conference on Hypertext and Social Media*, pp.85-94,2020.

[17] H. Yan, E. E. Fitzsimmons-Craft, M. Goodman, M. Krauss, S. Das, and C. Rehg , "Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention," *International Journal of Eating Disorders*, vol.52, pp.1150-1156,2019.

[18] S. Paul, J. S. Kalyani, and T. Basu , "Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks," *Working Notes of CLEF*, pp.1742-1749,2018.

[19] Y.-H. Huang, Y.-L. Lin, and L.-W. Ku , "A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text," *Proceedings of the IEEE*, pp.1742-1749,2018.

[20] K. Hemmatirad, H. Bagherzadeh, E. Fazl-Ersi, and A. Vahedian , "Detection of Mental Illness Risk on Social Media through Multi-level SVMs," *IEEE Access*, vol.8, pp.185028-185040,2020.

[21] M. P. Jain, S. Sribash Dasmohapatra, and S. Correia , "Mental Health State Detection Using Open CV and Sentimental Analysis," *Proceedings of the International Conference on Information Systems Security*, pp.465-470,2020.

[22] I. Syarif, N. Ningtias, and T. Badriyah , "Study on Mental Disorder Detection via Social Media Mining," *Proceedings of the International Conference on Information Systems Security*, pp.1-6,2019.

[23] K. R. Saradha, R. S. M., R. Saranya, and S. R. , "Sentimental Analysis to Detect Mental Illness," *Proceedings of the IEEE International Carnahan Conference on Security Technology*, pp.1031-1034,2022.

[24] C. Zucco, B. Calabrese, and M. Cannataro , "Sentiment analysis and affective computing for depression monitoring," *IEEE BIBM*, pp.1988-1995,2017.

[25] H. Zogan, I. Ahmed, S. S. Hu, and G. G. Xu , "A Multimodal Deep Learning Approach for Mental Disorder Prediction," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.1245-1254,2019.

[26] T. Zhang, A. M. Smith, and J. R. Brown , "Machine learning for mental health detection via longitudinal linguistic markers," *Scientific Reports*, vol.15(1), pp.1-12,2025.

[27] K. Pradeepa, S. J. Ram, and M. V. , "Mental Health Social Media Analysis Using Hybrid Deep Learning Models," *Advances in Intelligent Systems and Computing*, vol.1364, pp.285-294,2021.

[28] S. Sahoo, A. K. Nayak, and S. P. , "Deep Learning-Based Classification of

Mental Disorders from Social Media Content," *Multimedia Tools and Applications*, vol.82, pp.15316-15338,2023.

[29] M. M. Tadesse, H. Lin, and B. X. , "Linguistic markers of mental health in social media: A deep learning approach," *Scientific Reports*, vol.10, 2020.

[30] A. Widiastuti, R. K. S., and M. F. , "Mental Health Text Classification Using FastText and CNN: A Comparative Study," *Kinetik*, vol.7(2), 2022.

[31] A. Abbou, S. Mazouzi, and S. L. G. G. J. J. , "Machine-Learning-Based Approaches for Post-Traumatic Stress Disorder Diagnosis Using Video and EEG Sensors: A Review," *IEEE Sensors Journal*, vol.23(20), pp.24135-24151,2023.

[32] S. Deepika and V. Deivamani , "Artificial Intelligence-based Mental Health Check-up System for Adults," *Microprocessors and Microsystems*, vol.94,2022.

[33] S. Gupta, S. Agarwal, and A. Kumar , "An Investigation on the Audio-Video Data Based Estimation of Emotion Regulation Difficulties and Their Association With Mental Disorders," *IEEE Access*, vol.11, pp.74324-74336,2023.

[34] A. Srinivas, G. S. Rao, and K. V. S. R. P. , "A multi-modal mental health detection system based on social media data," *Computer Methods and Programs in Biomedicine*, vol.10, 2022.

[35] O. Hadjri, J. M. Bonaldi, and P. R. , "Automatic detection of schizophrenia: A systematic review," *L'Encéphale*, vol.44(6), pp.534-542,2018.

[36] A. Haque, S. A. G., and J. K. , "A Multimodal Approach to Mental Health Detection Using Deep Learning," *Computer Methods Programs inBiomedicine*, vol.245, 2024.

[37] B. Duan, S. L. Goodrich, and J. S. Huang , "An Integrative Approach for Subtyping Mental Disorders Using Multimodal Data," *Biological Psychiatry*, vol.95(10), pp.124-132,2024.

[38] K. G. Creswell, M. J. Sliwinski, and S. K. Bogucki, "Multimodal Mental Health Monitoring Using Passive Sensing and Machine Learning," *Sensors*, vol.19(17), 2019.

[39] S. C. Guntuku, R. K. Ganesan, and A. B. Ungar , "Multimodal Mental Health Analysis in Social Media: Exploiting Big Data for Depression Detection,"

Computer Methods and Programs in Biomedicine, vol.12,2022.

[40] R. Kshirsagar, J. O. Park, and T. S. Sethi , "CURE: Context- and Uncertainty-Aware Mental Disorder Detection," *CEUR Workshop Proceedings*,pp.156,2024.

[41] A. S. Uban, D. S. Chulvi, and P. Rosso (2021). "Predicting Mental Health Outcomes from Social Media Behavior Using Deep Learning," *Scientific Reports*, 11, 15124.

[42] Y. Zhu, L. W. Wang, and J. M. Zhang , "Multimodal Deep Learning with Attention Mechanisms for Mental Health Evaluation," *IEEE ICME*,2024.

[43] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol.10(1), pp.1-13,2018.

[44] A. Mandrekar, et al., "Multimodal Deep Learning Architecture for Mental Health Anomaly Detection Using Textual and Visual Expression Fusion," MDPI Sensors, 2024.

[45] B. Grimm, P. Yilmam, B. Talbot, and L. Larsen, "Multimodal Machine Learning for Video-Based Single Question Mental Health Assessment," npj Digital Medicine, 2025.

[46] H. Wang, T. Xu, and J. M. Zhang, "A Joint Late-Fusion Classification Framework for Clinician-Assisted Mood Disorder Evaluation Using BERT and Spatial Action Unit Encoders," IEEE Transactions on Affective Computing, 2024.

[47] X. Zhu, T. Gedeon, S. Caldwell, and R. Jones , "Detecting emotional reactions to videos of depression," *IEEE IV*, pp.198-203,2019.

[48] Y. Inoue, M. Eguchi, T. Kakinuma, and A. Ogami ,"Predicting physical and mental health status through interview-based evaluation of work stress," *Industrial Health*,vol.62(4), pp.144-156,2024.

[49] R. A. J. de Belen, V. Eapen, T. Bednarz, and A. Sowmya , "Using visual attention estimation on videos for automated prediction of autism spectrum disorder and symptom severity in preschool children," *PLOS ONE*,vol.19(2),2024.

[50] J. A. Naslund, A. Bondre, J. Torous, et al. , "Social media and mental health: Benefits, risks, and opportunities for research and practice," *Journal*

of Medical Internet Research, vol.22,2020.

[51] I. Barnett, J. Torous, P. Staples, et al. (2018). “Relapse prediction in schizophrenia through digital phenotyping: A pilot study,” *Neuropsychopharmacology*, 43, 1660–1666.

[52] D. S. Lehrer and J. Lorenz ,. “Anosognosia in schizophrenia: Hidden in plain sight,” *Innovations in Clinical Neuroscience*, vol.11, pp.10-17,2014.

[53] N. Kamal and M. Khan ,. “Linguistic markers of mental health disorders on social media using XGBoost,” *Journal of Computational Psychology*, vol.14, pp.112-125,2021.

[54] J. Suler, “The online disinhibition effect,” *CyberPsychology & Behavior*, vol.7, pp.321-326,2004.

[55] M. De Choudhury, S. Counts, and E. Horvitz , “Social media as a measurement tool of depression in populations,” *Proc. ACM Web Science*, pp.47-56,2013.

[56] Z. Zhang, L. Wang, and Y. Liu , “Multimodal sensing for depression risk detection: Integrating audio, video, and text data,” *Sensors*, vol.24, pp.1530-1545,2024.

[57] M. Alcañiz, J. Marín-Morales, and I. Abad , “Visual attention estimation for autism detection in children using computer vision,” *Journal of Clinical Medicine*, vol.11, pp.345-358,2022.

[58] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark , “Learning transferable visual models from natural language supervision,” *International Conference on Machine Learning*, pp.8748-8763,2021.

[59] Y. Li, Y. Xu, X. Mao, and C. Lu , “Symmetry and group in attribute-object compositions,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,2020.

[60] T. Nagarajan and K. Grauman , “Attributes as operators: factorizing unseen attribute-object compositions,” *ECCV*,2018.

[61] D. Huynh and E. Elhamifar , "Compositional zero-shot learning via fine-grained dense feature composition," *Advances in Neural Information Processing Systems*,2020.

[62] M. Lewis, N. V. Nayak, P. Yu, Q. Yu, J. Merullo, S. H. Bach, and E. Pavlick. "Does CLIP bind concepts? Probing compositionality in large image models," *Findings of the Association for Computational Linguistics: EACL* , pp.1487-1500 ,2024.

[63] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna "Crepe: Can vision language foundation models reason compositionally?," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,2023.

[64] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou "When and why vision-language models behave like bags-of-words, and what to do about it?," *International Conference on Learning Representations*,2023.

[65] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu. "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol.25, pp.281-304, 2022.

[66] A. Helmy, R. Nassar, and N. Ramadan. "Depression detection for Twitter users using sentiment analysis in English and Arabic tweets," *Journal of King Saud University - Computer and Information Sciences*, vol.36, pp.101-115 ,2024.

[67] S. Ji, T. Zhang, S. Ananiadou, and J. Tsujii. "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports*, vol.10 ,2020.

[68] M. A. Al-Garadi, M. M. De Choudhury, A. M. Hasan, et al. "Depression detection in social media posts using transformer-based models and auxiliary features," *Journal of Biomedical Informatics*, vol.145 ,2023.

[69] A. Waqas, A. A. Naqvi, A. M. Mubeen, et al. "Machine learning models to detect anxiety and depression through social media: A scoping review," *Computer Methods and Programs in Biomedicine Update*, vol.2 ,2022.

[70] H. AlSagri and M. Ykhlef. "Machine learning-based approach for depression detection in Twitter using content and activity features," *IEICE Transactions on Information and Systems*, vol.E103.D(8), pp.1825-1832,2020.

[71] J. Cha, S. Kim, and E. Park. "A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community," *Humanities and Social Sciences Communications*, vol.9, pp.1-12 ,2022.

[72] A. Radfar, S. Qorbani, M. Ghaffari, et al. "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Scientific Reports*, vol.15, 2025.

[73] S. Gaur, U. Kursuncu, A. Alambo, et al. ""Let me tell you about your mental health!" Contextualized classification of reddit posts to DSM-5 for web-based intervention," *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp.753-762,2018.

[74] Y. Lin, H. Lin, M. M. Tadesse, et al. "Multimodal machine learning for mental disorder detection: A scoping review," *Procedia Computer Science*, vol.225, pp.2929-2936 ,2023.

[75] K. Creswell, M. J. Sliwinski, S. K. Bogucki, et al. "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," *Sensors*, vol.24(2) ,2024.

[76] J. Hengle, A. Hengle, and S. Hengle. "Still depressing: A benchmark for depression-stress classification from online posts," *arXiv* ,2024.

[77] X. Liu, Y. Zhang, Z. Wang, et al. "It hears, it sees too: Multi-modal LLM for depression detection by integrating visual understanding into audio language models," *arXiv preprint arXiv:2511.19877* ,2025.

[78] J. Wang, Y. Zhang, H. Liu, et al. "Multimodal deep learning for automatic mental health assessment: A comprehensive survey," *Information Fusion*, vol. 96, pp. 102–124, 2023.

[79] Y. Li, X. Chen, J. Wu, et al., "Artificial intelligence for early mental disorder detection using multimodal behavioral data: A review," *JBI*, vol. 148, 2024.