Integrative Multimodal Analysis for Schizophrenia Cause Identification

A major project report submitted in partial fulfilment of the requirement for the award of degree of

Bachelor of Technology

in

Computer Science & Engineering

Submitted by

Sarthak Kurothe (211420)

Shivansh Kushwaha (211308)

Under the guidance & supervision of

Dr. Kushal Kanwar

Assistant Professor (SG)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, Solan - 173234 (India)

May 2025

Supervisor's Certificate

This is to certify that the major project report entitled "**Integrative Multimodal Analysis For Schizophrenia Cause Identification**" in partial fulfilment of the requirements for the award of the degree **Bachelor of Technology** in **Computer Science & Engineering**, in the Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is a bonafide project work carried out under my supervision during the period from July 2024 to May 2025.

I have personally supervised the research work and confirm that it meets the standards required for submission. The project work has been conducted in accordance with ethical guidelines, and the matter embodied in the report has not been submitted elsewhere for the award of any other degree or diploma.

Kushal Kanwar

(Supervisor Signature)

Supervisor Name: Dr. Kushal Kanwar

Designation: Assistant Professor (SG)

Department: Dept. of CSE & IT

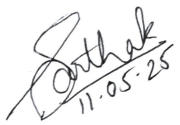
Date: 8th May 2025

Place: JUIT, Waknaghat

Candidate's Declaration

We hereby declare that the work presented in this major report entitled **'Integrative Multimodal Analysis for Schizophrenia Cause Identification'**, submitted in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** submitted in the Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from July 2024 to May 2025 under the supervision of **Dr. Kushal Kanwar,** Assistant Professor (SG), Department of Computer Science & Engineering and Information Technology.

We further declare that the matter embodied in this report has not been submitted for the award of any other degree or diploma at any other university or institution.



(Student Signature)
Name: Sarthak Kurothe
Roll No.: 211420
Date: 8th May 2025

(Student Signature)
Name: Shivansh Kushwaha
Roll No.: 211308
Date: 8th May 2025

This is to certify that the above statement made by the candidates is true to the best of my knowledge.

(Supervisor Signature)
Supervisor Name: Dr. Kushal Kanwar
Designation: Assistant Professor (SG)
Department: Dept. of CSE & I

Date: 8th May 2025
Place: JUIT, Waknaghat

ACKNOWLEDGEMENT

With immense gratitude, we extend our heartfelt thanks to the Almighty for his divine blessings, which have illuminated our path and enabled us to successfully complete the project – ***Integrative Multimodal Analysis for Schizophrenia Identification***. Our sincere appreciation goes to our esteemed Supervisor, ***Dr. Kushal Kanwar***, Assistant Professor (SG) in the Department of Computer Science & Engineering and Information Technology at Jaypee University of Information Technology, Waknaghat. Dr. Kushal Kanwar's profound expertise in the realms of Data Analytics, Machine Learning, Deep Learning, and programming languages has been instrumental in guiding us through this project. We are deeply indebted for his tireless support, patient mentorship, constructive criticism, and unwavering encouragement.

We would also like to express our gratitude to Dr. Kushal Kanwar from the Department of Computer Science & Engineering and Information Technology for his valuable assistance, which significantly contributed to the successful conclusion of our project.

Our sincere thanks extend to all individuals, whether directly or indirectly, who played a role in the triumph of this project. We acknowledge the support of the entire staff, both teaching and non-teaching, whose timely assistance and facilitation greatly aided our endeavour.

In closing, we wish to recognize and appreciate the enduring support and patience of our parents. Their unwavering encouragement has been a source of strength throughout this journey.

With gratitude,

Sarthak Kurothe

(211420)

Shivansh Kushwaha

(211308)

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviations	Meaning
fMRI	Functional Magnetic Resonance Imaging
EEG	Electroencephalography
fNIRS	Functional Near-Infrared Spectroscopy
F1-Score	Harmonic Mean of Precision and Recall
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
NLP	Natural Language Processing
ResNet	Residual Neural Network
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
ROI	Region Of Interest
SGD	Stochastic Gradient Descent
SSD	SSingle Shot Detector
AI	Artificial Intelligence
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
MFCC	Mel Frequency Cepstral Coefficients
TPU	Tensor Processing Unit
VGG	Visual Geometry Group
fNIRS	Functional Near-Infrared Spectroscopy

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Schizophrenia, a mentally challenged disease, can sometimes need massive examination of behavioural, linguistic, and physiological cues for proper diagnosis. The conventional diagnostic method is highly dependent on clinician's subjective observations, which are difficult to obtain at a handy pace and may vary over time. The goal of this project is to create an automated system based on multimodal analysis in order to support schizophrenia identification.

The approach includes analysis of data from three specific modalities: video, audio, and transcripts obtained during patient interviews, and video frames are analysed through the use of sophisticated computer vision techniques to obtain facial expressions and body pose features. Audio recordings are analysed with a view to extracting acoustic features such as pitch changes, pauses and spectral features. The NLP of text transcripts extracts linguistic features such as sentiment polarity, word counts and coherence.

These modalities are pre-processed in a mono-way manner and their properties are saved in structured form that is the basis for the integration into a consolidated scheme. The system is engineered to employ one-shot learning techniques, and therefore, it is very effective, even with a small pool of interviews. This report explains the preprocessing stage and identifies the solution in integrating the modalities into one consistent data set for training the model.

The proposed system can contribute to improved diagnostic accuracy, low occupation of clinicians, and can be functioned as a decision-support tool in mental health care. Future work will focus on multimodal fusion, model development, and deployment for real-world applications.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Mental health disorders are a major challenge to the health systems globally with as one of the most difficult and disabling conditions, schizophrenia remains. Schizophrenia, which at its worst takes 1% of the total world population, is defined by delusion, hallucination, disorganized thinking, social withdrawal, et cetera. The condition has consequences for both the patients' quality of life and the inordinate burden placed on families, caregivers, and healthcare professionals. Early detection and intervention is critical to preventing long-term effects from arising from the disorder, however current diagnostic methods are quite subjective and time consuming and are filled with more variability than not.

Current clinical techniques of identifying schizophrenia, in a traditional manner, heavily depend on clinical interviews and subjective observation of behavioral and linguistic cues. These assessments are carried out by experts in mental health professionals who assess patient's responses during the interviews. Although these means may be useful, their objectivity and consistency is not always perfect and particularly not from practitioner to practitioner and cross culturally. Further, stigmatization related to mental health disorders plays an active role in delaying diagnosis, which in turn will further the patient's condition. There is an urgent need for automated systems that could support clinical reviews by supplying objective data driven knowledge.

The sprouting AI and ML technologies have created new opportunities in diagnostics to mental health. It is possible to use multimodal data (video, audio, and text) and to analyze a variety of behavioral and physiological cues that may serve as indicators to schizophrenia. Multimodal analysis integrates several sources of data to give a better concept of a patient's situation. E.g. video data may reveal gaze or facial expression behavior or body language data and audio data can capture prosody of speech or pitch, or pauses. Text data emerging out of transcripts can provide linguistic and semantic clues, e.g., word frequency, sentiment, and coherence.

The purpose of this project is to create an automated system of schizophrenia detection based on the analysis of multimodal data. The system receives three modalities of data. video frames, audio recordings, and transcripts of text (for audio or video, transcribed manually into text). Features relevant for schizophrenia detection are extracted from each modality prior to preprocessing. For video data, face landmarks, body pose, and micro expression (highlighted in Fig. 2) are extracted using computer vision algorithms. Audio data is examined to recognize acoustic features such as pitch, intensity and spectral characteristics. An NLP algorithm is used to process the transcripts, in order to extract linguistic features including word count, sentence complexity, and polarity of sentiment.

A major challenge in undertaking this project is alignment and integration of data from these three modalities. The modalities do so at different timescales and may address unique aspects of the interview. Video frames, for example, are usually processed at a high frequency (30 frames/s), and audio features extracted over a fixed window (1 second). On the other hand, transcript characteristics are bound to spoken words and may not align directly with the temporal resolution of video or audio although there can be a portion of unified data which is common for both voice and text. To overcome this problem, careful synchronization and preprocessing should be performed to obtain a unified dataset from features of all modalities.

The project uses one-shot learning approaches that are particularly applicable to low data situations. Such one-shot learning models as Siamese networks and prototypical networks are developed on the principle that there is a possibility of developing appropriate representations for even a small number of training samples. This is vital for this project because the dataset consists of interviews from only 11 participants. Owing to the small size of the dataset yet one-shot learning techniques can generalize well by employing the power of the rich feature sets extracted from multimodal data.

The preprocessing pipeline in this project has several stages. Video frames extracted from the interview videos are processed using the Mediapipe thereby obtaining facial landmarks and pose information. Audio recordings are processed with Librosa to obtain pitch modulation and pauses, and the audio is decomposed into its spectral components. TextBlob and NLTK are used to break the transcripts into words and sentences while the sentiment polarity, subjectivity, and the coherence are extracted from the transcripts. These features are stored in structured forms (e.g., JSON file) in the form of each modality, and hence can be easily integrated and analyzed.

The combination of such modalities is a key step in the project. A single data structure that aggregates features from video, audio and text of each interview is developed. This aggregate representation sits as an input for the machine learning models. The project also has a mechanism of selective processing of certain videos to make computation efficient and scalable. For example, scripts have also been created to process only selected videos (e.g. interviews 10 and 11) during the debugging and testing process.

The applications of this project do not even limit to failed detection of schizophrenia. The techniques for multimodal analysis generated here can be used with other mental health conditions (depression, anxiety, and bipolar) as well. Using factual, data-based information, this system can be an aid to clinical assessments, decrease diagnostic discrepancies and enhance the quality of patients' care. In addition, the project is making a significant contribution to the expanding domain of AI in mental health by making an illustration of combining the integrative modal data for diagnostic purpose.

Finally, this project is, undoubtedly, a step towards the development of automated systems in mental health diagnostics. Through the use of video, audio, and text data, the proposed system tries to undertake a comprehensive and objective review of schizophrenia associated behavioural and linguistic markers. Multimodal analysis and one-shot learning techniques provide a strong framework for battling the problems of shortages of data and diagnostic variants. Future work will be focused on the optimization of integration of modalities, the training of machine learning models and validation of the system on the larger data sets. Using this work, we would like to push the state of mental diagnostics and support improved results for persons suffering from schizophrenia.

1.2 PROBLEM STATEMENT

Schizophrenia is an extreme and long-term mental health disorder which disrupts the thinking, feeling, and behavior of the person. It frequently leads to delusions, hallucination and disorganized speech, diminished emotional expression, poor cognition. Based on world health data, schizophrenia affects some 20-million people in the world, which concerns being a major health problem. The fact that it is a serious and common condition does not simplify the diagnosis of the disease.

Challenges in Schizophrenia Diagnosis

Schizophrenia is mainly diagnosed through clinical interviews and observation of behaviour together with self-reported symptoms. Clinicians use standardized diagnostic tools, such as the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) or ICD-10 (International Classification of Diseases), to identify patterns of behavior and symptoms. However, this approach has several inherent limitations:

1. Subjectivity and Variability:

• Diagnosis is usually subjective and requires the clinician's base it on his or experience, and his or her interpretation of the patient's response and actions.

• There is, moreover, an effect of cultural and linguistic differences on the interpretation of symptoms, hence variations in diagnosis among practitioners and regions.

2. Delayed Diagnosis:

• Schizophrenia develops insidiously, and there are mild or ambiguous symptoms in the first stages. This leads to delayed recognition and diagnosis, which can make the prognose poor in the long term.

• Stigma of mental health produces behavioural inhibiting individuals from accessing medical intervention in due course of time.

3. Comorbidities and Overlapping Symptoms:

• Certain symptomatic features of Schizophrenia are also seen in other psychiatric diseases, for example, bipolar disorder and Major depressive disorder. Especially without objective diagnostic tools, it is hard to distinguish these conditions.

• Substance abuse or anxiety as comorbid conditions, are likely to make the process of diagnosis more challenging.

4. Resource Constraints:

• Access to trained mental health practitioners is limited in low-resource settings, leading to underdiagnosis or misdiagnosis.

• Even in high-resourced healthcare systems, demand for mental health services frequently outweighs the supply of available clinicians.

5. Lack of Objective Biomarkers:

• In contrast to physical illnesses, where diagnosis is facilitated by biomarkers like blood tests or imaging studies, schizophrenia does not have clearly defined objective biomarkers. The majority of diagnostic attempts are based on subjective measures of behaviour and speech.

Potential of Multimodal Data in Mental Health Diagnostics

Technological progress and data science have made new possibilities available to tackle these problems. In particular, the combination of multimodal data—unifying video, audio, and text data—is a promising method for constructing objective and automatic schizophrenia diagnosis tools.

1. Video Data:

• Facial expressions, eye gaze, and body language are the most important nonverbal cues that reveal valuable information about an individual's mental state. For instance, reduced facial expressivity (blunted affect) and unusual patterns of gaze are characteristic of patients with schizophrenia.

• Latest computer vision methods, including pose estimation and facial landmark detection, allow one to extract those features from video recordings of interviews with patients.

2. Audio Data:

• Speech is a rich channel of information regarding mental health. Pitches, intonations, speech rates, and pausing can signal states of cognition and emotion.

• Speech abnormalities related to schizophrenic patients include Speech monotonicity, increased pauses, and lower fluency of speech. These features can be extracted by means of audio processing tools and managed quantitatively.

3. Text Data:

• The spoken or written language content and structure indicate a person's thought processes. Coherence, repetition, and strange word coupling are linguistic indicators associated with schizophrenia.

• The use of NLP techniques makes it possible to extract sentiment, complexity, and word frequency among other linguistic characters of the source from transcripts of patient interviews.

Through the use of these modalities, big data, a multimodal analysis system can acquire comprehensive picture of the behavioral, linguistic, and emotional characteristics of a patient.

Current Gaps in Research and Practice

Though there is the potential for multimodal data, there are huge gaps that must be filled to make this approach possible and efficient:

1. Integration of Modalities:

• Each mode (video, audio, text) is different and complementary information. However, combining these modalities into a single system still proves to be a technical and computational challenge.

• Effective integration call for the temporal alignment of modalities (e.g., matching audio characteristics with associated video frames) and combining has features that preserve individual contributions.

2. Limited Data Availability:

• It is not unusual that mental health datasets are relatively small because they are touchy data and difficulty of earning patient consent. This limits the ability

traditional machine learning models that would need significant volumes of datasets to perform well.

• Such methods as one-shot learning can provide possible solutions for the models to be trained with limited data, but they are still not yet popular in mental health research.

3. Ethical and Privacy Concerns:

• Ethical issues arise out of collecting and analyzing patient data privacy and consent. It is the guarantee of anonymity and security of patient data, important in winning trust and acceptance of such systems.

4. Validation and Generalizability:

• A lot of multimodal analysis apply to mental health on small, homogeneous datasets. Such data trained models may not well in general in populations or real-world clinical settings.

This project attempts to resolve these challenges by creating a system for schizophrenia detection that utilizes multimodality data analysis. The diagnosing of schizophrenia involves great difficulties subjectivity, in particular delayed recognition, and lack of objective biomarkers. Multimodal data analysis offers a promising solution that is driven by video, audio, and text to offer a holistic assessment of a patient's condition. This project contributes to filling in some of the most glaring gaps in the field by building a system that merges multimodal features even with limited data training models and holds performance in a structured way. By pumping more into the application of AI and data science, in mental health, this work seeks to increase diagnostic precision, increase patient outcomes, and support the overall mission of making easy access to and equal care for mental health possible.

1.3 OBJECTIVES

Schizophrenia is a complicated mental health condition which usually demands a deep evaluation of behavioral, linguistic, and physiological cue assessment for correct diagnosis. This project utilizes multimodal data, which include video, audio and video transcripts from patient's interviews, to create an automated system that presumes to

increase the objectivity and efficiency of schizophrenia detection. The goals of the project are as follows:

1. Process video frames to extract facial landmarks and body pose features using methods of computer vision, and record acoustic values such as pitch, pauses, and spectral properties from audio recordings. Transcripts will be processed in order to extract some linguistic features such as word-count, coherence, and sentiment.

2. Combine and harmonize features from the three modalities into one data so that there is temporal alignment between the video frames, audio segments, and interviews transcripts for each interview.

3. Train machine learning models, such as Siamese or Prototypical Networks, employing one-shot learning techniques to effectively recognize schizophrenia-related patterns despite limited data availability.

4. Measure the performance of the system using metrics such as accuracy, precision, recall, and F1-score, ensuring robustness through cross-validation.

5. Create a scalable and deployable system that will help clinicians to objective, data-driven, insights on the schizophrenia diagnosis, with the potential to be used in the practical world among clinical settings.

1.4 SIGNIFICANCE AND MOTIVATION OF THE PROJECT WORK

Schizophrenia is a big and powerful mental disease that affects millions of individuals around the world and has a much debilitating impact on their cognitive, emotional, and social functioning. It is a chronic condition that includes symptoms – delusions, hallucinations, disorganized thinking, and lack of emotional expression. Although it is a serious disorder, its diagnosis continues to be one of the major concerns because of the dependence on subjective clinical observations; the heterogeneity of the manifestation of the symptoms; and lack of objective biomarkers. Such limitations make it hard to achieve early detection, consistent diagnosis, and timely. Intervention, highlighting in urgency the need for new ideas for mental health diagnostics.

This project is of great importance, in tackling these challenges, through the implementation of analysis of multimodal data in order to produce a thorough and unbiased approach to

schizophrenia diagnosis. The classical forms of diagnosis, which are very dependent on clinician expertise and patient-reported symptoms are however commonly subjective and subject to variability. By this project, it provides a data- driven alternative that is driven and can pick up subtle markers of schizophrenia, for example facial expressions, speech pattern and linguistic coherence. The ability to combine this conversion into a single system of modalities is likely to greatly improve diagnostic accuracy, particularly in picking up on the softer early-stage symptoms that might easily slip past during manual assessments.

One of the main driving forces behind this work is the absence of objective biomarkers for mental illnesses. Compared to physical illness that must be assessed with blood work or imaging, mental illness is harder to reach. Research can validate diagnoses but schizophrenia is totally reliant on observation of behavior and descriptions of symptoms. The gap identified in the information above opens the door to potential of artificial intelligence and machine learning how to transform the diagnostic tasks of mental health into a quantitative, reproducible insights. The use of multimodal data is not only an addition of objectivity to the process, but it also enables the system identifying presumably patterns and correlations across modalities to be rather holistic.

The development in machine learning and deep learning technology again enhances the motivation of this project. Technologies like Mediapipe for video processing, Librosa for audio processing, and natural language processing libraries for text processing enabled it to extract sophisticated features from every modality. When combined together, these features form a powerful dataset to train sophisticated models. Besides, utilization of one-shot learning methods like Siamese networks and Prototypical networks mitigates the overall issue of small sample sizes in mental health studies. There are limited interviews to draw from, thus these methods provide the ability to create good models that are able to generalize well using small sample sizes.

Another primary motivator is the potential that such a system could enhance accessibility and scalability in mental healthcare. Countries, especially in low resource settings, are experiencing the acute inadequacy of trained mental health practitioners, a system that automatically could assist with such information might bridge the gap between clinicians, by providing vetted, impartial information. It would enable more consistent and accurate diagnoses with less reliance on highly trained experts. Moreover, by automating some

aspects of the diagnostic process, this system can alleviate the burden on clinicians and leave them with the responsibility of treatment and patient care.

Its social influence is the second driving force. Mental illness tends to be something of a stigmata, and this causes delayed diagnosis and treatment. By providing a scientific, objective framework for diagnosis, this project is able to dispel some of the mystique surrounding mental illness and promote earlier intervention. It is simpler to believe in a diagnostic system based on measurable data than on open interpretation. In the long run, this can result in better drug treatment adherence and quality of life for patients with schizophrenia.

This work is also inspired by the more far-reaching implications for research in mental health. The methodologies and frameworks thus developed in this project can be extended to other psychiatric symptoms such as depression, anxiety, and disorders like bipolar disorder. With the demonstration of the feasibility and efficiency of multimodal analysis, the present project lays the foundation for future work. in using AI to addressing the mental health issues. It also offers to the growing amount of evidence supporting combined application of technology in medicine, which lays out the opportunities based on evidence-driven solutions to complement and enhance tradition practice.

This project is a major leap forward in the area of mental health diagnostic providing an alternative, revolutionary method of approaching the intricacies of the diagnosis of schizophrenia. It aims to do so by integrating multimodal data with high-level algorithms surpass the constraints of traditional diagnostic processes and be a scalable and objective, and accessible solution. The impetus for this work is based on its reality for making a. effective difference in the patient's and clinician's lives, which will make way for the future in which mental diagnoses are both more effective and fairer.

1.5 ORGANIZATION OF PROJECT REPORT

This report outlines the detailed analysis and development of a multimodal system for detecting schizophrenia via video, audio and text data. The objective of this project is to lead the reader via methods, findings, and contributions, with a clear understanding of the innovative ways in which these modalities have been combined and used to manage the problems of mental health diagnostics.

CHAPTER 1: INTRODUCTION

The first chapter is about the concept of employing multimodal data to aid in schizophrenia detection. It provides the context and significance of this work in the area with a focus on the limitations of the conventional model of diagnosis and promise of machine-based, data-driven solutions. The chapter defines the objectives and scope of the project, Highlighting the conjoining of video, audio, and text information; it paves the ground for the ensuing chapters by defining the relevance and the impact this suggested system would probably have.

CHAPTER 2: LITERATURE REVIEW

Chapter 2 reviews already published work on the detection of schizophrenia with attention to that aspect of its application using machine-learning and multimodal analysis in mental-health. It reviews recent developments in audio-based speech processing, video-based behavioral analysis, and linguistic feature (of web content) extraction from text data. The chapter informs us strengths and weakness of these techniques, barreling down the challenges of aligning and blending a range of modalities. Gaps existing methods are emphasized; the necessity of a single, multimodal system is warranted as envisioned in this project.

CHAPTER 3: SYSTEM DEVELOPMENT

This chapter presents the methodology followed to create the proposed system for schizophrenia detection. It describes pre-processing of each modality: facial landmarks and body pose extraction out of video frames and utilization of acoustic features in audio files and research into a linguistic nature of transcripts. The chapter also describes how these modalities are included in an integrated data structure and how one-shot learning is used to train models and the technical decision with which they were made, including the tool and algorithm selections.

CHAPTER 4: TESTING

Chapter 4 is the system implementation and testing phase. It outlines the process of tagging video, audio and transcript information to ensure synchronization for precision feature representation. In the chapter, the training of a machine learning model is described using processed multimodal dataset and the validation parameters utilized for testing of that system is able to detect schizophrenic patterns effectively.

CHAPTER 5: EVALUATION AND RESULTS

In this chapter, the experimental results are given showing the performance of the system in identifying schizophrenia markers. Accuracies, precisions, recalls and F1-score, etc. are given to measure the effectiveness of the integrated multimodal system. The results are contrasted with traditional solutions to expose the advantages of using video, audio, and text data. The chapter concludes with the real-world implications of the system for mental health diagnostics.

CHAPTER 6: CONCLUSION AND FUTURE SCOPE

The final chapter is a conclusion of the project's most important results, remarks on the issued recommendations needing to be implemented in schizophrenia diagnosis with multimodal analysis and troubles faced while working on the project and measures taken. The chapter finishes by peeking into the future directions for the system, fine-tuning the system with advanced models, expanding the dataset for improved generalization, and discussion of use of this the approach to the other mental health conditions.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW OF RELEVANT LITERATURE

Gowtham Premananth, Yashish M.Siriwardena, Philip Resnik, Sonia Bansal, Deanna L.Kelly, Carol Espy-Wilson et al. [1] proposes a multimodal approach for symptom measurement of schizophrenia through an integration of brain imaging, neurocognitive examination, and analysis of social behavior. The strategy emphasizes the fact that the multiple must be integrated. Data modalities generation with the aim to obtain a greater understanding of schizophrenia spectrum disorders. The approach aims at enhancing symptom severity assessment through integration several data sources. However, one of the main challenges to the interpretation is suggested by the study the interaction effects of multiple modalities. Despite the possible advantages of multimodal approaches, the level of complexity in integrating these types of data, typically leads to the synthesis being difficult, and thus, can, in turn, be a hindrance to the overall interpretability of the findings. The study lays basis for future work that would improve future synthesis of these modalities and overcome the interpretability issues of them.

Göker, H. [2] is dedicated to the application of one-dimensional convolutional neural networks (CNNs) for the Schizophrenia detection from EEG signals automatically. This study proves that 1D-CNN can be very efficient in diagnosing schizophrenia by processing EEG, proving the most significant benefit of this method is offered by the fact that it is able to capture temporal patterns of EEG signals, which are crucial towards diagnosis of psychiatric disorders. Nevertheless, the process is limited to EEG data and thus diminishes its ability of generalization to other types of multi-modal data, such as text and video. While the CNN parameterization process is effective in EEG-based detection, future work may be forced to explore hybrid models with other modalities that will ensure maximum accuracy of the diagnosis and robustness of the diagnostic systems.

Siuly, S., Guo, Y., Alcin, O.F. et al. [3] investigated the application of deep residual networks (ResNets) towards the automatic detection of schizophrenia based on EEG recordings. Residual networks are a category of deep learning models recognized for their proficiency in learning intricate features more effectively using skip connections that assist

in preventing vanishing gradients when trained. The authors discovered that ResNets enhanced the recovery of feature-relevant information from EEG signals to improve classification performance for schizophrenia detection. But one of the biggest drawbacks to this method is the requirement of large amounts of training data. The model must have a great deal of diversified data in order to perform very well, something that might not be feasible within the realm of schizophrenia given that there are so few annotated data sets available. This makes techniques that can successfully operate on lower data sets or use data augmentation methods especially necessary.

Chuang CY, Lin YT, Liu CC, Lee LE, Chang HY, Liu AS, Hung SH, Fu LC.[4] provided a multimodal method of evaluating schizophrenia symptom severity by focusing on linguistic, acoustic, and visual cues. This approach points at the possibility of combining the two sources of data to enrich the overall picture of schizophrenia. The study demonstrated that each of the modalities – linguistic cues drawn from speech, acoustic markers, such as intonation, pauses and visual cues such as face expression and gesture – may give vital insights into the severity of symptoms. Nonetheless, one of the most critical challenges in this approach is the synchronizing these modalities, which are not always perfectly matched with each other as regard timing or frequency. All these challenges notwithstanding, the work illustrates the power of multimodal analysis, and recommends for further improvement in data synchronization and fusion techniques may better the reliability and accuracy in your symptom evaluations.

A. Kanyal, S. Kandula, V. Calhoun and D. H. Ye [5] utilized the fusion of functional MRI (fMRI), structural MRI (sMRI), and single nucleotide polymorphism (SNP) information for the classification of schizophrenia. Integrating imaging data into genetic data with multi-modal deep learning, the research observed multimodal fusion performing better than individual-modality techniques with regard to classification accuracy. The study emphasized the need for incorporating neuroimaging and genomic data to enhance diagnostic accuracy in schizophrenia. Yet, one of the main challenges that the study has found is standardizing MRI data between datasets. Differences in imaging protocols and methods may introduce biases that can influence the performance and generalizability of the model. The study indicates that standardizing MRI data might result in stronger and more reproducible results for schizophrenia classification.

Chilla, G.S., Yeow, L.Y., Chew, Q.H. [6] used ensemble machine learning methods to classify schizophrenic patients and healthy controls according to neuroanatomical markers (cortical volumes, brain areas). The study recorded classification accuracies over 83% to 87%, sensitivity being between 90% to 98%. This strategy shows the efficacy of neuroanatomical markers for differentiating schizophrenia, which fits the emerging pool of research that supports the use of abnormalities of the brain structure as diagnostic markers. Nevertheless, the authors report that the neuroanatomical markers examined were few, and future investigation might increase the number of features in the set to enhance performance of classification. Furthermore, although the results are encouraging, the employment of ensemble methods does require users to devote special attention to the problem of computational complexity as well as a potential problem of overfitting when working with datasets of small size.

Das et al. [7] examined the use of multimodal fusion with explainable AI (XAI) to detect schizophrenia. The combination of the fMRI and sMRI data demonstrated that multimodality fusion was able to yield greater accurate and interpretable results relative to single modality techniques. In this context, the XAI approach proves to be particularly useful since, with the aid of this approach, clinicians can learn the rationale for model's predictions – something that is important for those applications to be used in the real world. However, the study does point to the need for additional refining of explainability methods, specifically clinical utility of these methods. Although XAI offers potential to improve the interpretability of the model, the complexity of method may limit its applicability in the larger clinical settings where transparency and ease of use are key.

Srivastava et al. [8] combined fMRI and EEG data using deep learning to enhance schizophrenia diagnosis accuracy. By combining these two modalities, it became possible to perform more robust FES, collecting both the patterns of functional connectivity in the brain, as well as temporal dynamics of the EEG signals. A comparison of results showed a significant increase in diagnosis accuracy by the use of multimodal data as compared to each modality separately. However, authors indicate that datasets need to be larger for the method to be cross-validated in different populations, and for generalizability. Furthermore integrating different modalities data presents challenges in terms alignment of time and processing which the research seeks to address through careful pre-processing and system design.

C. -R. Phang, F. Noman, H. Hussain, C. -M. Ting and H. Ombao [9] have suggested CNN-based approach for schizophrenia detection based on EEG pattern concentrating on connectome-based features. Gathering the results of the study, by examining brain connectivity based on EEG signals, better classification performance was achieved, which is the evidence of how connectivity patterns can act as biomarkers for schizophrenia. The connectome-based approach demonstrates intricate relations among various brain regions that frequently change in psychiatric disorders such as schizophrenia. However, the authors propose that more testing is warranted on larger sets of data for improving the model's generalization and providing its reliability in other patient cohorts. This work highlights the need for brain connectivity analysis in schizophrenia detection while being hampered by smallness and diversity of data set.

Palani Thanaraj Krishnan, Alex Noel Joseph Raj, Parvathavarthini Balasubramanian, Yuanzhu Chen [10] applied Empirical Mode Decomposition (EMD) and entropy measures for interpreting multichannel EEG signals for the application in schizophrenia detection. EMD breaks down complex signals into intrinsic mode functions thus the signal can be decomposed to make features extraction easier on the EEG data. The entropy-based measurements in this study were used to determine the dimensions of complexity and regularity of EEG signals, which are transformed in schizophrenic patients. This approach worked out well for detecting schizophrenia; however, the work was confined to EEG whereas other modalities such as fMRI or text were not included. The use of multimodal data may increase robustness of the detection system and also make it more applicable to clinical practice.

David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Jonathan Edward Robinson, Sridha Sridharan, Patrick J Johnston, Kristin R Laurens, Clinton Fookes [11] discussed early detection of schizophrenia in children through development of deep learning techniques for EEG responses. The study showed that the deep learning models could successfully detect at risk children at high precision and recall. Early diagnosis is critical in schizophrenia because early treatment has a higher likelihood of achieving successful long-term all-over results. However, the authors point out that their results should be confirmed with a wider age range and various types of data, for example, video or audio. The attention to EEG only may prevent the effective extrapolation of the model, and combining other modalities may offer a more complete picture of the schizophrenia risk.

Desai, R., Porob, P., Rebelo [12] have applied the Wavelet Packet Transform (WPT) and Gaussian Process Classifier (GPC) for classifying EEG data to detect schizophrenia. The results from the combination of WPT and GPC for feature extraction and grouping inhibition of EEG signals proved excellent to the analysis of EEG signals.

Tikka SK, Singh BK, Nizamie SH, Garg S, Mandal S, Thakur K, Singh LK [13] employed high-density EEG data were used in conjunction with SVM classifiers in schizophrenia diagnosis. It was shown; this approach resulted in high classification accuracy of this approach; this attests to the potential of EEG to detect neural patterns involved in schizophrenia. The ability to obtain higher density EEG enables a deeper insight into the brain activity, which is vital for the specification of schizophrenic patient's and healthy control's brain activities. However, the study reveals some limitations like high computational complexity of processing the high-density EEG data. Moreover, overfitting is vulnerable to the model, especially in application with small samples, a phenomenon common in psychiatric research that is constrained by the availability of small sets of data. The necessity of bigger datasets and more efficient computational procedures are emphasized to guarantee reliability and generalizability of the model.

Ji N, Liang Ma, Hui Dong, and Xuejun Zhang [14], studied the combination of Discrete Wavelet Transform (DWT) and Empirical mode decomposition (EMD) with Approximate entropy for the extraction of features from EEG signals for the detection of schizophrenia. It is observed that the combination of the DWT and EMD methods substantially improved the feature extraction that resulted in better representation of the characteristics of the EEG signal which are related to schizophrenia. The addition of approximate entropy, a technique, which intends to capture the complexity and cyclicality in a signal, supplied some further information regarding the cognitive and emotional states of patients with schizophrenia. Despite these positive results, the authors reported that this approach is only applicable to EEG data and cannot apply other modalities such as MRI and fMRI to provide complementary information for a holistic diagnosis. Further research is needed to integrate these methods with imaging modalities to improve the accuracy of detection and extend the usefulness of the procedure.

Sugai Liang, Yinfei Li, Zhong Zhang, Xiangzhen Kong, Qiang Wang, Wei Deng, Xiaojing Li, Liansheng Zhao, Mingli Li, Yajing Meng, Feng Huang, Xiaohong Ma, Xin-min Li, Andrew J Greenshaw, Junming Shao, Tao Li [15] published a paper about

semantic knowledge transfer that described the study aimed at classification of first-episode schizophrenia based on multimodal brain features such as structural and diffusion MRI. The study was able to have high classification accuracy in identifying schizophrenia individuals through the consolidation of MRI-based features that address both the structure and connectivity of the brain and therefore, a much better understanding of schizophrenia. This approach is especially useful in detecting early-stage schizophrenia in which there are developing structural changes in the brain. Multimodal imaging is perceived as a step forward in increasing diagnostic accuracy of schizophrenia. However, the study concentrated on first episode schizophrenia only and did not extrapolate to later stages of the illness. Such a limitation implies further investigation to determine whether the same approach can be generalized to higher stages of schizophrenia or whether other features/modalities are necessary for diagnosis of later phases.

R. Salvador [16] studied the combination of several MRI modalities (functional structural and diffusion imaging) to increase the diagnostic possibilities of schizophrenia. The study demonstrated an improvement in diagnostic accuracy by combining various types of MRI data thereby indicating complementariness in different imaging techniques. Functional MRI (fMRI) renders information about the brain activity, structural MRI (sMRI) image brain anatomy, while diffusion MRI (dMRI) shows white matter integrity. Using the combination of these modalities provides a more precise account of the structure and function of the brain in schizophrenia patients, which, in turn, enhances the diagnostic outcomes. Nevertheless, the research also specified that integration of the modalities also involves an amplified level of computational complexity, thus making the process more resource consuming. This may limit the wide use of such methods in the clinical environment, where efficiency and cost-effectiveness matters. The study proposes that more work remains to be done on streamlining the process of integration, and to make multimodal MRI-based approaches more practical for routine clinical use.

Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM [17] examined the use of social media behavioral markers to detect schizophrenia using the machine learning research in combination with clinical appraisals. This cutting-edge research looked at possible indicators to track online including changes in language use, social activity online, and amount of emotional expression and was able to detect potential markers of schizophrenia. The adoption of machine learning made it possible to implement extraction

of responsible features from social media data based on which the feature analysis of clinical appraisals created a new dimension for detecting mental health disorders. However, the findings need additional validation using larger samples and across different platforms, according to the study. The dependence on social media as a source of data has the potential to present both opportunities and challenges, such as ethical issues and privacy issues, which must be considered prior to implementation such techniques into practice in the clinical practice. With all the deficiencies, though, the study exposes fascinating possibilities for the use of online behavior as an additional tool for diagnosing schizophrenia, and specifically, real-time.

Elzbieta Olejarczyk, Wojciech Jernajczyk [18] utilized graph-based analysis of brain connectivity in schizophrenia, on the basis of EEG signals. The study showed that there are significant differences in brain activity between schizophrenia patients and healthy controls, which focused on the importance of brain network disruptions during the pathophysiology of schizophrenia. Through graph theory, the authors could present the brain as a network of connected regions and, by doing so, locate abnormalities of connectivity patterns that are frequently connected to the disorder. Although useful insights about the neurodynamic underlying schizophrenia can be derived from the study, the small sample size used in the study had limitations in terms of the generality of the findings. The authors propose that further studies of larger clinical datasets would aid in establishing robustness of these findings and would enable biomarkers based on brain connectivity patterns to be developed. The use of graph theory on EEG data constitutes a novel strategy towards understanding schizophrenia and it offers promise in terms of improving the diagnostic accuracy.

Varshney, A., Prakash, C., Mittal, N., Singh, P. [19] has been as concerned as he was with trying a multimodal approach that combines functional MRI (fMRI) and structural MRI (sMRI) data for the diagnosis of schizophrenia. The study found that the combination of the two types of MRI, made it possible to obtain more accurate diagnostic tool while detecting both activities of the functional brain and the structural changes affecting schizophrenias. The research showed that combining sMRI with fMRI information would have a highly significant overall impact on diagnostic accuracy when compared to utilizing either modality on its own. But the research has been constrained by focusing on MRI data only without integrating other perhaps beneficial modalities such as EEG or genetic data. This drawback indicates that there is a need for further study of combining more modalities, including

genetic or behavioral measurements, with the current approach to increase diagnostic performance. Schizophrenia diagnosis between multimodal data is an encouraging opportunity for elaborating more profound diagnostic tools which call for deeper investigation and cross-sample validation in the patients of different types.

Yunchao Yin, Jianting Cao, Qiwei Shi, Danilo P. Mandic, Toshihisa Tanaka, Rubin Wang [20] proposed to use Multivariate Empirical Mode Decomposition (MEMD) for processing of EEG signals in the problem of quasi brain death including schizophrenia. The study shows that MEMD was capable of decomposing EEG signals to obtain meaningful features associated with brain activity in critical states. The authors discovered patterns, related to the changed brain function by analyzing the energy distribution of EEG-signal, which may be important for understanding schizophrenia. Although this approach yielded positive results, the study mainly concentrated in critical brain states such as brain death, and was not expanded to cover schizophrenia diagnosis in general. This limitation narrows the application of the method to routine schizophrenia screening. However, the study underlines the promise of MEMD for analyzing the complex EEG data and implies that more work is necessary in order to examine the use of MEMD in combination with other modalities, including fMRI, or behavioral data to form a more comprehensive diagnostic tool for schizophrenia.

Table 2.1: Literature Review Table

S. No.	Author & Paper Title [Citation]	Journal/ Conference (Year)	Tools/ Techniques/ Dataset	Key Findings/ Results	Limitations/ Gaps Identified
1.	Gowtham Premananth, Yashish M.Siriwardena, Philip Resnik, Sonia Bansal, Deanna L.Kelly, Carol Espy-Wilson A Multimodal Framework for the Assessment of the Schizophrenia Spectrum [1]	Interspeech (2024)	Brain imaging, neurocognitive assessments, social behaviour analysis	Integration of various data modalities to assess symptom severity	Difficult to interpret interaction effects between modalities
2.	Göker, H. 1D-convolutional neural network approach and feature extraction methods for automatic detection of schizophrenia.[2]	SpringerLink (2023)	1-D CNN, EEG Signals	Accurate detection using 1D-CNN for EEG signal processing	Limited to one data type, may not generalize to other modalities

3.	Siuly, S., Guo, Y., Alcin, O.F. et al. Exploring deep residual network based features for automatic schizophrenia detection from EEG[3]	SpringerLink (2023)	Deep Residual Networks, EEG	Improved feature extraction using residual networks	Needs extensive training data for higher accuracy
4.	Chuang CY, Lin YT, Liu CC, Lee LE, Chang HY, Liu AS, Hung SH, Fu LC. Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic and Visual Cues.[4]	IEEE Trans Neural Syst Rehabil Eng (2023)	Linguistic, acoustic, and visual data, multimodal analysis	Holistic assessment of symptom severity from multiple data sources	Challenges in synchronizing multimodal data.
5.	A. Kanyal, S. Kandula, V. Calhoun and D. H. Ye Multi-modal deep learning from imaging genomic data for schizophrenia classification.[5]	IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) (2023)	MRI, SNP, 1-D CNN, FBIRN, BSNIP, COBRE Datasets	Multi-modal fusion outperforms single modality in SZ classification, leveraging sMRI, fMRI, and SNPs for better accuracy.	Requires standardization of MRI data across datasets to avoid biases due to different imaging techniques.

6.	Chilla, G.S., Yeow, L.Y., Chew, Q.H. Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods.[6]	Scientific Reports (2022)	Ensemble methods, neuroanatomical markers (cortical volumes, areas, etc.)	Achieved classification accuracies (83-87%), sensitivity (90-98%), and correlations with QoL scores.	Limited range of neuroanatomical markers to explore further.
7.	Das A, et al., An explainable AI approach for schizophrenia detection using multimodal fusion[7]	IEEE Access (2022)	fMRI, sMRI, Explainable AI	Multimodal fusion combined with explainable AI provided interpretable results for SZ diagnosis.	Explainability methods need refinement for broader clinical applications.
8.	Srivastava P, et al., Multimodal fMRI and EEG deep learning analysis for schizophrenia diagnosis[8]	Neurocomputing (2021)	fMRI, EEG, Deep Learning	Integrating fMRI and EEG for multimodal analysis significantly improved SZ diagnosis accuracy.	Requires larger datasets for cross-validation across modalities.
9.	C. -R. Phang, F. Noman, H. Hussain, C. -M. Ting and H. Ombao A multi-domain connectome CNN for identifying schizophrenia from EEG patterns.[9]	IEEE Journal of Biomedical and Health Informatics (2020)	EEG, CNN	Improved SZ diagnosis using connectome-based EEG features with CNN for enhanced classification performance.	Requires additional testing on larger datasets to improve generalization.

10.	Palani Thanaraj Krishnan, Alex Noel Joseph Raj Schizophrenia detection using Multivariate Empirical Mode Decomposition and entropy measures from multichannel EEG signal.[10]	Biocybernetics and Biomedical Engineering (2020)	EEG, Empirical Mode Decomposition (EMD), Entropy	Effective use of EMD and entropy measures in EEG signals for SZ detection.	Method limited to EEG data, lacking multi-modal comparisons.
11.	David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Jonathan Edward Robinson Identification of children at risk of schizophrenia via deep learning on EEG Responses[11]	IEEE Journal of Biomedical and Health Informatics (2020)	EEG, RNN,CNN	Early SZ detection using deep learning on EEG patterns in children, with high precision and recall.	Requires validation on broader age groups and data types.
12.	Desai, R., Porob, P., Rebelo EEG Data Classification for Mental State Analysis Using Wavelet Packet Transform and Gaussian Process Classifier[12]	Wireless Personal Communications (2020)	EEG, Wavelet Packet Transform (WPT), Gaussian Process Classifier (GPC)	WPT and GPC demonstrated high effectiveness in classifying EEG data for SZ diagnosis.	Limited focus on EEG modality without considering other data sources. .

13.	Tikka SK, Singh BK, Nizamie SH, Garg S, Mandal S, Thakur K, Singh LK. AI-based classification of schizophrenia using EEG and SVM[13]	Indian Journal of Psychiatry (2020)	EEG, Support Vector Machine (SVM)	High-density EEG and SVM successfully classified SZ patients with high accuracy.	High computational complexity and potential overfitting with small sample sizes.
14.	Ji N,,Liang Ma ,Hui Dong and Xuejun Zhang EEG feature extraction using DWT and EMD combined with approximate entropy[14]	Brain Sciences (2019)	EEG, Discrete Wavelet Transform (DWT), Empirical Mode Decomposition (EMD), Approximate Entropy	Combined DWT and EMD methods showed improved feature extraction for SZ detection in EEG data.	Further work needed to combine these with other modalities like MRI or fMRI.
15.	Sugai Liang, Yinfei Li, Zhong Zhang, Xiangzhen Kong, Qiang Wang, Wei Deng, Xiaojing Li, Liansheng Zhao, Mingli Li, Yajing Meng Classification of First-Episode Schizophrenia Using Multimodal Brain Features[15]	Schizophrenia Bulletin (2019)	Structural and diffusion imaging, brain MRI features	High classification accuracy using multimodal imaging	Focused on early-stage schizophrenia, lacks generalization to other stages
16.	R. Salvador Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia[16]	Frontiers in Neuroscience (2019)	MRI modalities (functional, structural, diffusion)	Improved diagnostic accuracy through multimodal integration	Complex model integration may increase computational cost

17.	Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals[17]	Journal Of Medical Internet Research (2017)	Machine Learning, Social Media Behavioral markers	Identified potential online markers for schizophrenia	Needs further validation with larger sample size and diverse platforms
18.	Elzbieta Olejarczyk , Wojciech Jernajczyk Graph-based analysis of brain connectivity in schizophrenia[18]	PLoS ONE (2017)	EEG, Graph Theory	Graph-based connectivity analysis revealed significant differences in SZ brain activity.	Limited sample size; requires larger clinical datasets.
19.	Varshney, A., Prakash, C., Mittal, N., Singh, P. (2016). A Multimodel Approach for Schizophrenia Diagnosis using fMRI and sMRI Dataset[19]	Intelligent Systems Technologies and Applications (2016)	fMRI, sMRI, multimodal analysis	Integration of functional and structural MRI for improved diagnosis	Limited to MRI data, lacks incorporation of other modalities .
20.	Yunchao Yin , Jianting Cao, Qiwei Shi , Danilo P. Mandic, Toshihisa Tanaka, and Rubin Wang Analyzing the EEG Energy of Quasi Brain Death using MEMD[20]	APSIPA ASC (2011)	EEG, Multivariate Empirical Mode Decomposition (MEMD)	MEMD effectively analyzed EEG signals for critical brain states, including schizophrenia.	Focus on critical brain states limits its applicability to wider SZ diagnosis.

2.2 KEY GAPS IN THE LITERATURE

Although a lot has been achieved in the detection and classification of schizophrenia using a multimodal approach, a number of important gaps are still there that need to be explored further. The integration of multimodal data is one of the major challenges that are clearly brought out in the literature. For example, it is difficult to interpret the interaction effects between video, audio, and text modalities in measuring the schizophrenia symptom severity; Premananth et al.[1] mentioned this. This complexity calls for superior techniques of data synchronization and fusion, which in most cases are under-researched to date.

The other gap that has been identified is the limited generalization of models towards varied datasets. Many researches, for example [2] Göker and [3] Siuly et al., demonstrate that although their models perform very well on certain data type, such as EEG signals, they fail on other modalities, or more-powerful datasets containing more varied information. According to Göker [2], 1D-CNNs for EEG are working well, but only for one type of data and Siuly et al [3] raise concerns on deep residual networks that need a lot of data to be trained in order to be more accurate, meaning that stronger models are necessary when it comes to different data sources and also smaller sample sizes.

The absence of standardized data sets is the other significant disadvantage. Research such as that by Kanyal et al. [5] indicate that it is possible to combine various sources of data, like MRI and SNPs' data in order to get more accurate results but the variety of quality and format of the dataset, especially in the case of the brain imaging data, can prevent training and evaluation of the model. Standardization across imaging techniques is imperative in order to eliminate biases, and make it possible to apply models in various clinical settings.

Furthermore, interpretable models, especially when using such complex forms of algorithms as deep learning, also remain a problem. According to Das et al, [7], the combination of multimodal data with explainable AI offers some level of transparency but additional improvement of the explainability approaches is needed for wider clinical application. This is especially important for clinicians who must recognize logical foundations for diagnoses produced by AI to be able to believe in and utilize these systems in reality.

Furthermore, a number of studies (Chuang et al.[4]; Varshney et al.[19]) stress the difficulty of generalizing findings from different stages of schizophrenia. For example, Varshney's study aimed at first-episode schizophrenia that does not give a clue as to subsequent stages of the illness disease, therefore, more research should be done on the stage specific diagnostic models based on the progression of the illness.

Finally, although in the individual studies promising findings were reported, there is a paucity of real-world validation as well as clinical integration. It is the most common research, such as Phang et al [9] and Srivastava et al [8] conducted and concluded, that Multimodal models perform well in environment-controlled or small-scale studies, but the scalability and practicality of these kinds of methods in clinical practice can be better confirmed through stronger testing. This difference between theoretical performance on a system and its performance in the real world is one that needs to be bridged in order for these systems to be actually useful.

Briefly, despite the potential shown in the literature for multimodal approaches towards schizophrenia detection, much remains to be done in how data is integrated, how generalization is possible in diverse datasets, how techniques are standardized, and interpretability and real-world applicability. It will be vital to fill the gaps to ensure the automated schizophrenia detection systems are not just theoretical advances but practical tools that can improve clinical outcome.

CHAPTER 3
SYSTEM DEVELOPMENT

3.1 REQUIREMENTS AND ANALYSIS

The construction of a multimodal system for detecting schizophrenia involved careful planning and the appropriate combination of hardware and software to facilitate video, audio and text processing. The task complexity required a strong setup for processing large data bases and computing challenging machine learning models. Below is the summary of the vital components that built the system below:

1. **Hardware:** That a massive data processing for this project needed considerable computational power. Key high-performance GPUs, NVIDIA Tesla V100, A100 or RTX 3090, were chosen based on their capabilities to speed up both training and inference steps of deep learning models. The GPU's supported efficient video frame, audio signal and transcript data processing. For larger workloads and to allow for flexibility cloud-based solutions such as AWS, Google and Azure were also integrated into the platform. This fusion saved us the trouble of having to address scalability and reliability separately since a balance was achieved within the hybrid setup for an easy performance of the resource intensive tasks.

2. **Software:** Multimodal data integration and deep learning were considerations behind the software environment. Python was adopted, primarily, because of its huge libraries for data science and machine learning. TensorFlow and PyTorch were the main frameworks used with the desired flexibility that came with training and fine-tuning complex models. The system was provided by pre-trained models for video processing, such as ResNet, and for text analyses including BERT, thus reducing development time. Additional tools including video processing specialized OpenCV and frameworks and libraries of NLTK were used to enable smooth experimentation provided a solid base for building and testing models.

3. **RAM:** To handle large volumes of data incorporated in this work, adequate memory was needed to prevent bottlenecks in processing. Not less than 16 GB of RAM was needed to run the extensive computations that were essential for training deep learning models. This meant that routines such as batch processing of data and execution of several tasks concurrently were all performed without cuts. Once the system had enough RAM, it became possible to perform complex calculations with ease which simplified the model development process.

4. **Storage:** To house datasets, model weights and training logs, the project needed a great storage capacity. Video frames, audio files and transcripts had to be loaded and retrieved at high speeds, for this reason SSDs were selected as opposed to the regular hard drives. The blessings of faster read/write speeds of SSDs brought down the latency and delivered efficient data fetching and processing. This simplified storage system was a necessary aspect to ensure smooth workflows when models were built and tested.

5. **Network:** Availability to the internet with a stable fast internet connection was extremely important because downloading of necessary datasets, pre-trained models and libraries were required. The project also used cloud-based computation resources, which needed a consistent and reliable connection. This provided for uninterrupted transfer of data and activities such as accessing huge repositories and uploading/downloads of files from the cloud undertaken in an efficient way. Real-time collaboration and remote computation within the development process was made possible with high-speed connectivity too.

6. **Development Environment:** Around Python was built the development environment that was supported by TensorFlow, PyTorch, OpenCV among other libraries for machine learning and multimodal data processing. Richly used during the bi-initial stages of the project for prototyping/testing configurations of models, Jupyter Notebooks offered a convenient platform for visualizing results interactively. In case of a larger scale of development and debugging, PyCharm was also used, having tidy and feature-rich environment. This combination of these tools made experimental work more efficient, which also streamlined the overall coding process.

7. **Scalability:** Scalability From the platform's conception, the system was designed in such a way that is scalable. Since the project was based on data integration modalities, the system had to be capable of accommodating possible extensions, for example, processing a larger dataset or addition of another modality. The cloud-based platforms gave the opportunity to scale the computational resources when needed, so that the system could be able to meet the increasing demands. The scalability allowed us to expand the features of the system for future applications including its deployment in real world clinical settings.

3.1.1 Analysis:

This project concerns developing a multimodal system for the identification of the presence of schizophrenia by combining video, audio, and text data. By critical investigation of the issue became evident, that conventional diagnostic approaches, which involve a significant degree of subjective clinician judgments, frequently fail at delivering reliable and repeatable outcomes. This brought to light the necessity of an objective, data-driven approach which will be able to take employ sophisticated machine learning algorithms to muddy the diagnostic process. Through the integration of several modalities, the system is to be able to capture the whole spectrum of behavioral linguistic and acoustic markers associated with schizophrenia.

A review of today's approaches identified a number of weaknesses. Models based on one type of data, for example EEG or MRI, frequently do not cover the whole spectrum of indicators typical of schizophrenia. Generalize these findings by single-modality systems tested on different datasets or in clinical contexts of daily reality is also fraught with difficulties for these single-modality systems. Even multimodal approaches, which hold great promise, fail with the complexity of aligning and integrating data from disparate sources – as well as the substantial computational power needed in these unifications. Such findings guided the choice of datasets, tools, and model architectures for the project.

To overcome these challenges the system was designed with attention to both hardware and software requirements in mind. The requirement to process big datasets and to train deep learning models heavily relied on the use of high-performance GPUs and cloud resources in order to ensure that the processing followed a path of efficiency and scalabilities. For purposes of the analysis on video data and for the purposes of text processing, pre trained models: such as ResNet and BERT, were also used in conjunction within strong frameworks

for audio feature extraction. These tools were selected not only on the basis of their performance but also on the flexibility of use that made experimentation and model optimization so easy.

The constraints of the traditional approaches for diagnosing schizophrenia played an important role in shaping the development of the project. These traditional techniques are typically unable to fuse disparate data sets thus curtailing their capacity to identify subtle multimodal signatures of the disorder. By filling the voids, this project's target is to deploy an integrated system that properly aligns and processes multimodal data. Such an approach is not only aimed at improving the diagnostic accuracy but also prepares an environment for the future use of such approaches in the clinical settings, giving a scalable and an efficient solution to one the most complex aspects of mental health diagnostics.

3.2 PROJECT DESIGN AND ARCHITECTURE

3.2.1 METHODOLOGY

The design of such a system for schizophrenia detection implements a rather holistic strategy, integrating high-level processing methods for video, audio and text data into a common analytical architecture. The system utilizes multimodal transformer models coupled with one-shot learning more effectively to carry out limited dataset analyses while capturing the complex relationships that exist concerning the three types of data modalities. This approach is designed to pave the way for smooth integration and synchronization of the various features collected from these varied depositories, ending up with an all-round determination of schizophrenia related markers.

Preprocessing of raw inputs is performed during data preprocessing. Video data, which is made of frames taken out from interview recordings, is fed using pre-trained models such as ResNet, in order to extract meaningful visual features such as facial expressions, movement of eye and body pose landmarks. For the case of audio recordings, the approaches of Mel Frequency Cepstral Coefficients (MFCC) and pitch variation analysis are used to draw out the speech-related acoustic features. At the same time, text transcripts of the interviews are tokenized and analyzed via natural language processing (NLP) approach in terms of aspects such as linguistic coherence, sentiment polarity, and semantic contrast. The unstructured data gets structured in such a way that is appropriate to use for more processing through these preprocessing stages.

The crux of the system is the multimodal transformer architecture in which elements from the three modalities are merged into a unified system. In the transformer model, embeddings from video, audio and text are processed where, with the help of self- attention mechanisms, the most relevant patterns and interactions are observed within the data types. This approach guarantees that important temporal and contextual relationships are preserved such that the system has the ability to pick subtle behavioral and linguistic markers that are typical of schizophrenia.

Having scarcity of labeled data in this domain, one-shot learning has been employed by the system in an attempt to address such a challenge. Models such as Siamese Networks, and Prototypical Networks are used to learn feature representations that are devastatingly effective, despite the small dataset size. By targeting the differences and similarities between data points; one-shot learning improves the system's capacity of generalizing well and predict effectively with minimum training examples.

The learning procedure is iterative and includes optimization of the model using appropriate functions of loss, contrastive for tasks of similarity, and cross-entropy for the purpose of solving classification problems. Performance of the system is measured by means of a number of key metrics such as precision, recall, F1-score and area under the receiver operating characteristic (ROC) curve. Such metrics give a clear picture of how good the system is in detecting schizophrenia across all modalities hence its reliability and robustness.

The final stage is detection and identification stage, where the system synthesizes outputs from the modality for a comprehensive diagnosis. Incorporating video, audio, and text data enables a deeper analysis than is possible in conventional single-modality methods. The efficiency of this rearranged multimodal system is demonstrated in terms of efficiency by comparison to unimodal systems, illustrating better capturing of schizophrenia multifacetedness.

This method is not only effective for identifying benefits of multimodal transformers and one-shot learning in processing scarce and varied data but it demonstrates the capabilities of the linkage between video, audio and text analysis to develop a more precise and credible diagnostic instrument. Through the narrowing down of the shortcomings of the conventional approaches, this system is meant to improve reference to the schizophrenia detection and the understanding of it thus making it useful for clinical applications.

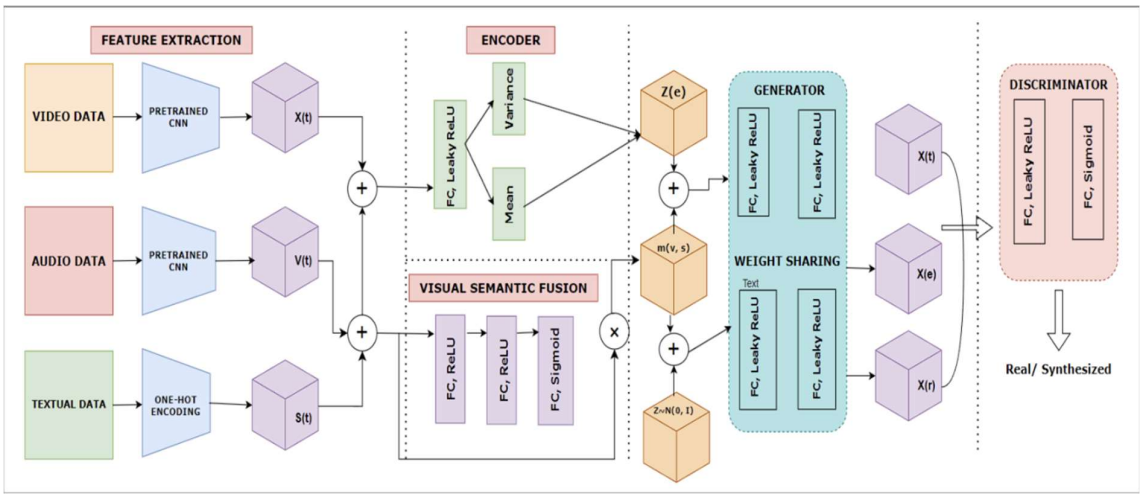


Fig. 3.1: Multimodal Variational Framework for Feature Fusion and Synthesis

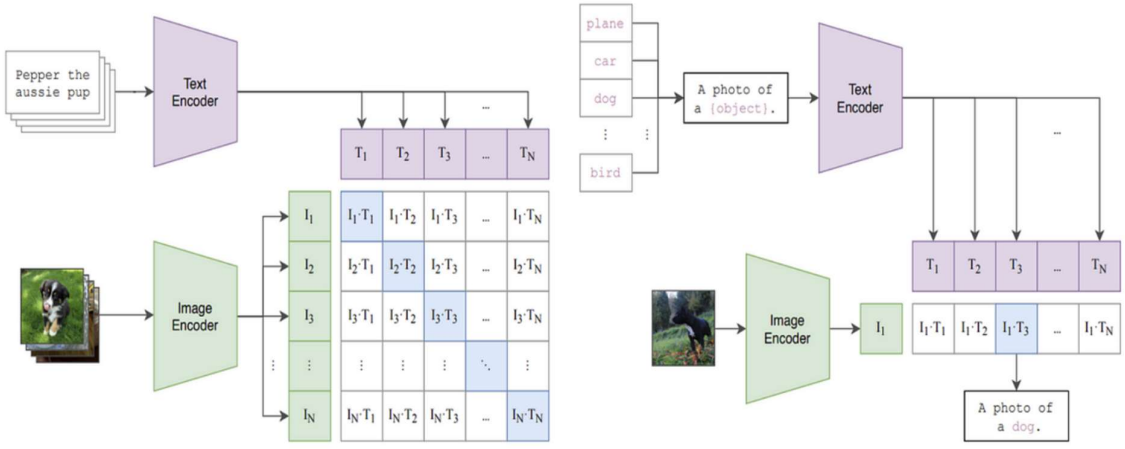


Fig. 3.2: Multimodal Learning Framework: Text-Image Embedding and Cross-Modal Representation

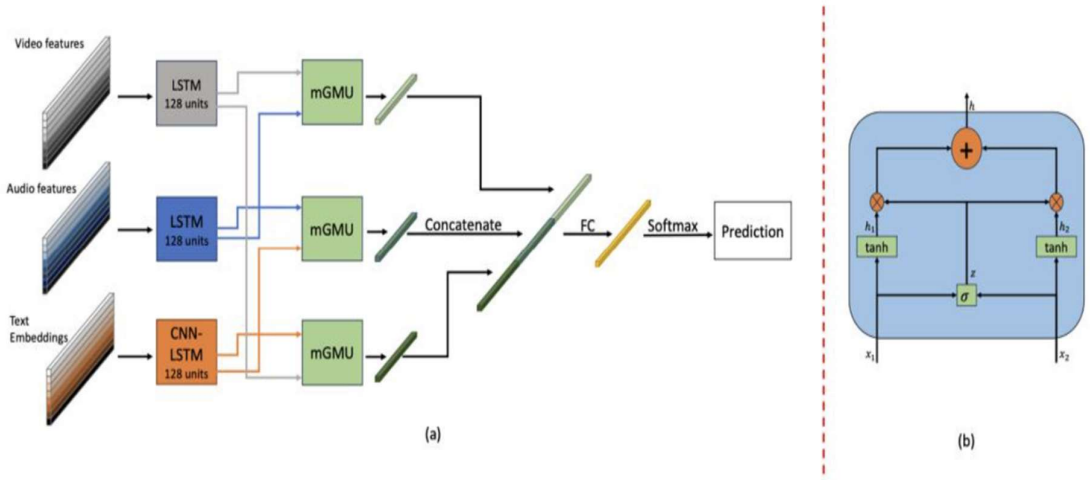


Fig. 3.3: Multimodal Fusion Architecture with mGMU and Feature Integration

3.2.2 DATA PREPARATION

In this project we created systematic procedure for management of multimodal data, such as video and audio and text, in order to effectively detect schizophrenia. Every type of data had their preprocess pipeline for it to filter through to pick the most meaningful features and ensure everything can interact easily within an integrated model. Computers vision methods were used for analysis of video frames in order to capture visual cues such as facial expressions and gestures. Audio recordings were transformed to extract characteristics of speech (tone and pauses) while text transcripts were improved using natural language processing techniques to bring out linguistic patterns. Such meticulous preparation made the data from the three sources accurate, align well, and suit integration into the system.

1. **Video Data Preparation:** The recorded video data includes respective recording of the interviews that each have been broken into individual frames to extract the visual features that pertain to schizophrenia related behaviours. Frames are analysed using state-of-the-art Computer vision tools such as Mediapipe to detect facial expression, micro-expression, body pose, gesture, and eye movement patterns. Such features are used to detect lack of facial expression, strange posturing or lack of eye contact among others. Processed frames are timestamped and saved into structured types including JSON files, which retain the temporal order with associated audio and text data for integration-based analysis.

2. **Audio Data Preparation:** From the audio recordings of the interviews, acoustic features are extracted to identify the speech-related pattern prevailing commonly in schizophrenia. Preprocessing aims at noise reduction, extracting the speaker's voice, feature extraction through the use of tools such as Librosa to obtain the computation of Mel-Frequency Cepstral Coefficients (MFCCs), pitch variations, speech rate and pauses. These characteristics are used to capture speech that was monotonic speech, speech with slowed processing, or disorganized. Projections of extracted features are used for numerical multivectors and correlated with video and transcript data for multimodal analysis.

3. **Text Data Preparation:** Text transcripts are tagged using Natural Language Processing methodologies of extracting linguistic and semantic features. The text consists of sentence tokenization and word tokens, and there are analyses that are carried out to estimate the text's sentiment level, coherence, and complexity level. Sentimentalized aspects of sentence length, word repetition, logical flow are extracted, to detect disorganized or fragmented thought patterns. The processed features in structured files are synchronized with audio and video modalities to offer an integrated representation of the patient's communication style.

4. **Synchronization Across Modalities:** To combine multimodal data, visuo-acoustic alignment is carried out by timestamping the video frames, audio segments and/or transcript text or setting predefined intervals. It maintains that it is possible to assign each frame with the appropriate audio and text data so that the dataset is unified for analysis. Temporal alignment is very important in maintaining the relationship between modalities to be able to analyse behavioural, acoustic and linguistic markers at once for specific detection of schizophrenia.

```python
# Process individual video frames from a .mp4 file
def process_video(video_path, output_dir):
    os.makedirs(output_dir, exist_ok=True)

    # Capture video
    cap = cv2.VideoCapture(video_path)
    frame_count = 0

    while cap.isOpened():
        ret, frame = cap.read()
        if not ret:
            break

        rgb_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
        face_results = face_mesh.process(rgb_frame)
        pose_results = pose.process(rgb_frame)

        # Extract features
        features = {"face_landmarks": [], "pose_landmarks": []}
        if face_results.multi_face_landmarks:
            for landmarks in face_results.multi_face_landmarks:
                features["face_landmarks"] = [
                    {"x": lm.x, "y": lm.y, "z": lm.z} for lm in landmarks.landmark
                ]

        if pose_results.pose_landmarks:
            features["pose_landmarks"] = [
                {"x": lm.x, "y": lm.y, "z": lm.z, "visibility": lm.visibility}
                for lm in pose_results.pose_landmarks.landmark
            ]

        # Save features for each frame
        output_file = os.path.join(output_dir, f"frame_{frame_count:04d}.json")
        with open(output_file, "w") as f:
            json.dump(features, f)

        frame_count += 1

    cap.release()

# Process all video segments
for i, video_segment in enumerate(video_segments_dir, 1):
    video_files = [f for f in os.listdir(video_segment) if f.endswith('.mp4')]
    output_segment_dir = os.path.join(results_dir, f'video_segments_{i}')
    os.makedirs(output_segment_dir, exist_ok=True)

    for video_file in video_files:
        video_path = os.path.join(video_segment, video_file)
        output_dir = os.path.join(output_segment_dir, os.path.splitext(video_file)[0])
        process_video(video_path, output_dir)
```

Fig. 3.4: Implementation of preprocessing the video frames

```python
import librosa
import numpy as np
import os
import json
from pydub import AudioSegment

# Define directories
audio_dir = "/content/drive/MyDrive/Datasetr/Dataset/Audio/"
processed_audio_dir = "/content/drive/MyDrive/Datasetr/processed_audio/"
os.makedirs(processed_audio_dir, exist_ok=True)

def process_audio(audio_path, output_path):
    # Convert .mp3 to .wav using pydub
    audio = AudioSegment.from_mp3(audio_path)
    temp_wav_path = audio_path.replace(".mp3", ".wav")
    audio.export(temp_wav_path, format="wav")

    # Load audio file
    y, sr = librosa.load(temp_wav_path, sr=None)

    # Extract features
    features = {
        "chroma_stft": librosa.feature.chroma_stft(y=y, sr=sr).mean(axis=1).tolist(),
        "rmse": float(librosa.feature.rms(y=y).mean()),
        "spectral_centroid": float(librosa.feature.spectral_centroid(y=y, sr=sr).mean()),
        "spectral_bandwidth": float(librosa.feature.spectral_bandwidth(y=y, sr=sr).mean()),
        "rolloff": float(librosa.feature.spectral_rolloff(y=y, sr=sr).mean()),
        "zero_crossing_rate": float(librosa.feature.zero_crossing_rate(y).mean()),
        "mfcc": librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13).mean(axis=1).tolist()
    }

    # Save features as JSON
    with open(output_path, "w") as f:
        json.dump(features, f)

    # Remove temporary .wav file
    os.remove(temp_wav_path)

# Process all audio files
for audio_file in os.listdir(audio_dir):
    if audio_file.endswith(".mp3"):
        audio_path = os.path.join(audio_dir, audio_file)
        output_path = os.path.join(processed_audio_dir, f"{os.path.splitext(audio_file)[0]}.json")
        process_audio(audio_path, output_path)
```

Fig. 3.5: Implementation of preprocessing the audio files

```
    # Tokenize words and sentences
    words = word_tokenize(text)
    sentences = sent_tokenize(text)

    # Extract features
    features = {
        "word_count": len(words),
        "sentence_count": len(sentences),
        "average_word_length": np.mean([len(word) for word in words]),
        "average_sentence_length": np.mean([len(sent.split()) for sent in sentences]),
        "sentiment_polarity": TextBlob(text).sentiment.polarity,
        "sentiment_subjectivity": TextBlob(text).sentiment.subjectivity
    }

    # Save features as JSON
    with open(output_path, "w") as f:
        json.dump(features, f)

# Alternative: Process transcript using spaCy (if NLTK fails)
import spacy
nlp = spacy.load("en_core_web_sm")

def process_transcript_spacy(transcript_path, output_path):
    with open(transcript_path, "r") as f:
        text = f.read()

    doc = nlp(text)
    words = [token.text for token in doc if token.is_alpha]
    sentences = list(doc.sents)

    features = {
        "word_count": len(words),
        "sentence_count": len(sentences),
        "average_word_length": sum(len(word) for word in words) / len(words),
        "average_sentence_length": sum(len(sent) for sent in sentences) / len(sentences),
        "sentiment_polarity": TextBlob(text).sentiment.polarity,
        "sentiment_subjectivity": TextBlob(text).sentiment.subjectivity
    }

    # Save features as JSON
    with open(output_path, "w") as f:
        json.dump(features, f)

# Process all transcripts
for transcript_file in os.listdir(transcripts_dir):
    if transcript_file.endswith(".txt"):
        transcript_path = os.path.join(transcripts_dir, transcript_file)
        output_path = os.path.join(processed_transcripts_dir, f"{os.path.splitext(transcript_file)[0]}.json")
        try:
            process_transcript_nltk(transcript_path, output_path)
        except Exception as e:
            print(f"NLTK failed for {transcript_file}: {e}")
            print(f"Falling back to spaCy for {transcript_file}")
            process_transcript_spacy(transcript_path, output_path)

print("Processing completed. JSON files are saved in:", processed_transcripts_dir)
```

Fig. 3.6: Implementation of preprocessing Transcripts

3.3 IMPLEMENTATION

The phase of implementation of this project is designed as a phase of structured and holistic procedure, but it has not been implemented yet. While the system design and data preparation are done, the following are the steps that require to be taken; building the functional system, which combines all parts, and testing it with the real-world scenarios.

The first stage of the implementation process will be a pipeline to process the multimodal inputs. Video data will use pre-trained ResNet models such as to do the processing in extracting the visual features such as facial expressions, body gestures, and eye-movements. Audio recordings will be processed with Librosa, to gather acoustic features such as MFCCs, pitch changes and silence during speech. Linguistic patterns, sentiment and coherence will be extracted from text data of transcripts using natural language processing techniques. These features will then be aligned and arranged into a unifying format to be used in the model.

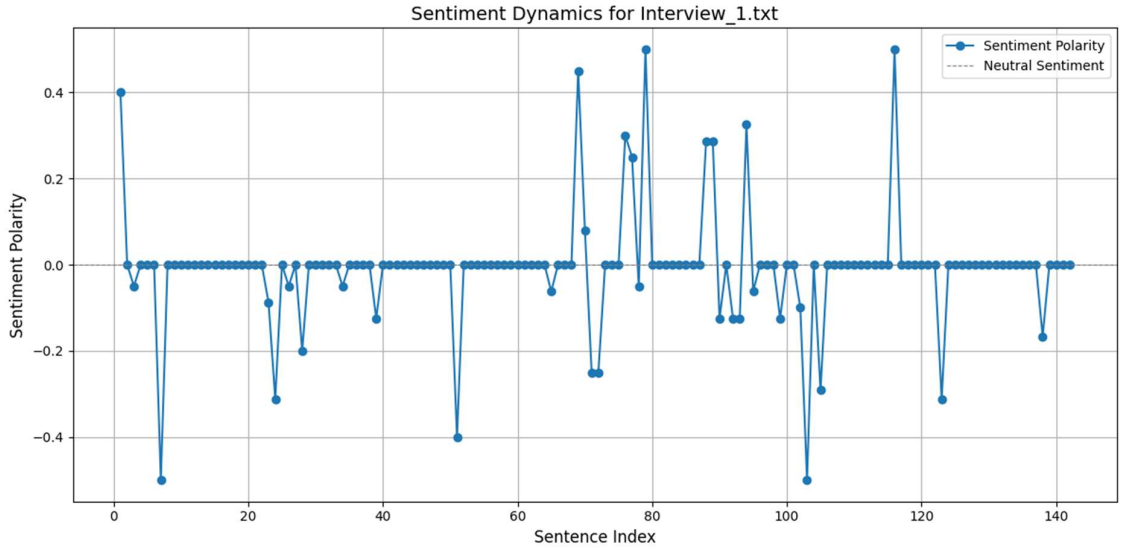


Fig. 3.7: Sentimental Dynamics of Interview_1

The centre of the system will be a multimodal transformer model, capable of combining features from multimodalities of video, audio and text. The transformer will apply self-attention mechanisms enabling it to observe interplay and relationships between the modalities, thus catching subtle patterns related to schizophrenia. To counter the problem of having little amount of data, the system will also make use of one shot learning techniques, including Siamese Networks, that will allow the system to train the model on small data sets without losing accuracy.

The performance of the system will be evaluated during the next stage of the project. Performance in the system detection of schizophrenia related markers will be evaluated by such measures as precision, recall, F1-score and ROC-AUC. In addition, unimodal approaches will be compared to show the benefits of using a multimodal framework.

Once developed the system will be tested and optimized heavily to make performance finer. Through these steps the model will be robust and will be ready for future clinical applications. The next phases will also be concerned with scalability, providing the system with a possibility to cope with bigger data sets and accommodate various real-life examples.

3.4 KEY CHALLENGES

The problem of developing a multimodal system for the detection of schizophrenia imposes a set of specific challenges in establishing a complex data mixing system but also in establishing the system as a practical tool on real-world terms.

1. **Aligning Mutimodal Data:** One of the grandest challenges involves gathering video, audio, and text data whole preserving their relationships. Each type of data is based on different timeframes and possesses different characteristics that do not perfectly fit together. For instance, for matching a particular video frame to its corresponding audio clip and transcript, there is need for great synchronization in order to have accurate analysis.

2. **Scarcity of Data**: Availability of large labeled datasets containing all three modalities is not plentiful. Privacy issues and the sensitive nature of mental health data limits this type of resources as well. Such lack of data makes it difficult to train machine learning models properly, which usually require large volume of information to get a hang of the patterns. The small dataset also makes it easier to overfit which presents challenge in application of the system to new cases.

3. **Diversity of Features**: Video, audio, and text provide their own distinctive (quantitative) descriptions, while not all extracted features are equally useful to diagnose schizophrenia. For example, making faces or gestures, depending on one's personality or cultural difference, or ways of talking, for example, tone or stop, can be driven by outside influences, say anxiety. The process of identifying those features

that are most relevant to the detection of schizophrenia is one that is not easily or unambiguously conducted and involves a complicated iterative process of judgement.

4. **High Computational Demands**: To process and analyze data from a combination of modalities, one must invest a lot of computational efforts. Multimodal transformers or one-shot learning architectures models may be resource demanding, especially on large data sets. Efficient training of these models, without losing precision is one of the challenges often necessitating GPUs or cloud platforms.

5. **Generalization to Real-World Data**: The other great challenge is another is that the system should learn to excel on other socio-demographic and unseen sets based across population. The variety of recording quality, the examinee's recording conditions, and the patient's demographics may all influence the accuracy of the model. The system must be resilient in terms of handling these differences to achieve consistency in results in diverse real-life situations.

6. **Ethical and Privacy Considerations**: The same applies to working with sensitive mental health data, there comes the question of ethics. Confidential, and protection of data by law is of utmost importance. To get the data for use for research purposes, one needs strong anonymization procedures, protection over data storage and ethical approval. These aspects need to be weighed against the requirement to build a meaningful and meaningful system.

CHAPTER 4

TESTING

4.1 TESTING STRATEGY

To make the multimodal schizophrenia detection one of the robust and effective tools, the testing strategy that would be comprehensive should be applied during the entire development process. There will be several phases under the strategy, for example, unit testing, integration testing, functional testing, edge-case testing, performance testing, and robustness testing to adequately test out the system's performance in various actual scenarios.

The first phase, unit testing, will be concerned with investigating the functionality of each separate part of the system. For this project, the units to be tested in this project are the video, audio, and text data preprocessing pipelines in addition to the feature extraction models for each modality. Here during this phase, we will verify that all of these individual components are functioning correctly – so anything from video features such as emotive deliveries; audio features such as enunciation of words or tonality; to text features such as sentiment analysis…all are being processed correctly. By testing each piece on its own, we can catch all kinds of problems early and correct them before you glue them all together into the greater system.

After unit testing, we will proceed with integration testing. At this stage the integration of the various elements in the system will be considered. We will see how the video, and audio, and the text features behave inside the multimodal transformer model. The aim is such that the features from different modality are well aligned and combined in a deserved manner such that the overall system makes accurate predictions. All the problems that may arise out of how the various modalities are synchronized, or incompatible with one another will be addressed in this stage.

Next, we will perform a functional test to confirm the system delivers its basic objectives. During this phase we will test if the system is compatible to detect and classify schizophrenia markers in this way across the video, audio, and text data. We will test the system's ability to address different real-world interview situations and evaluate how it performs reliably well in standard situations. Such phase is vital to confirm that the system can identify the

schizophrenia behaviors and speech patterns, including disorganized speech promising unusual facial expression.

To test further how the system treats unexpected or challenging situations, the system is supposed to be tested to read edge-cases. During this stage the system will have to live with bad inputs including poor video, distortion of audio or defectives text. By the “worst-case” checking of the system, we will be able to observe its performance when dealing with unclear or noisy data. This aids to see the parts where the system may find it hard and would help us make necessary changes to be able to deal with such situations at the future.

Performance testing will determine how fast and efficiently data are processed in the system. We will time how long the system takes to make predictions and its capability of dealing with large datasets in a reasonable amount of time. This is important particularly to real time applications where speed is essential. We will also test how scalable our system is by running the system with various hardware configurations to build realism on how well the system deals with larger datasets or more complicated parameters when deployed to the real-world – proving that it is scalable.

Ongoing testing will verify the system’s reaction in managing error or missing data of different kinds in the model. Then we are going to check and validate the system using noisy, missing, or damaged data to ensure that it does not fail or produce the wrong outputs. This is necessary to ensure the system is dependable in real-life changing scenarios where occasionally the data quality is not going to be optimal.

At last, we will have to conduct accuracy validation test, to validate the overall performance of the model. We will measure various factors like precision, recall and F1-scores of the system to see how well is the detection of schizophrenia-related pointers. With this benchmark data with various interview scenarios, we will fine-tune the model on these outputs to enhance the performance.

4.1.1 PROGRAMMING LANGUAGE:

For this project, we have selected Python as our programming language of choice due to the ease with which it can be read, the flexibility of the language and popularity of its use in areas of such as data science and machine learning. Python’s clean syntax simplifies the code, which makes the code base easily developed and fast to run. Due to its dynamic typing, then

we don't need to declare variable types beforehand, this makes for faster development. Python also supports various programming styles from which one can get to resolve any kind of problem; they include the object-oriented programming and the functional programming among other.

The great thing about Python is the multitude of libraries and frameworks (which Python developers love). For our project we'll have to use these libraries in order to do the majority of the video processing, audio analysis, and natural language processing. The richness of tools and libraries ecosystem around Python allows us to build the model without reinventing the wheel. Plus, it is very easy to receive support from the large developer community, so now it is an ideal solution for the complicated tasks we engage in.

4.1.2 AI LIBRARIES/FRAMEWORKS

1. **TensorFlow:** It will support us to train and optimize our models. It presents solid support for pre-trained models and can handle the intense computational tasks necessary to process large data sets fairly and efficiently.
2. **PyTorch:** It is the most efficient library for building and testing the models. Its versatility, and ease of use makes this library ideal for the project especially when it comes to integrating video, audio and text data.
3. **OpenCV:** It is used for processing and analysis of video data. It is a great tool for such tasks as facial expression recognition and pose estimation, that are necessary for interpreting visual cues of the videos.
4. **Scikit-learn:** This is used for assessing the models and carrying out activities such as classification and clustering so that we are able to determine how well our system is performing.
5. **Keras:** It is an easy-to-use library based on TensorFlow, to be able to create and test various deep learning models quickly.
6. **NumPy:** NumPy is a crucial library in dealing with numerical operations, particularly when dealing with a lot of video, audio, and text data.

4.1.3 GOOGLE COLAB AS IDE

For this project, we are using Google Colab as our primary environment for code and since it has powerful cloud capabilities and free access to GPUs and such is important for training

and running a deep learning model. This is particularly useful for such things as working over big datasets and running complicated models such as multimodal transformers. Colab's integration with Google drive is superb and therefore one can easily manage and share files, which are great for a collaboration. The platform also allows the use of requisite Python libraries (TensorFlow, PyTorch, OpenCV and Keras) for implementing the models in this project. The well-designed interactive Jupyter notebook interface in Colab enables bundling and combining code, visualizations and explanations all in one place so it is much easier to play, track how we go, and improve models as we go. Overall, Colab provides a flexible, quite friendly environment in which development and collaborative work for this project are simplified.

4.2 TEST CASES AND OUTCOMES

As we proceed with the creation of the multimodal schizophrenia detection system, we will conduct a wide variety of significant tests to guarantee that all happens according to the plan. First, unit testing will validate that each of these data pipelines; video, audio and the text generates the features that we are looking for. For example, we will verify that video can identify facial expressions properly, audio can determine speech patterns and text can pick up the linguistic features correctly. If something goes wrong in these steps we will fix it immediately.

Then there will be integration testing to make sure the system will be able to integrate capabilities across all three modalities, video, audio, text into a single input. This step will then ensure that the system is processing the data both right and effectively. After that we will perform functional testing which will test the ability of the system to identify the markers of schizophrenia such as disorganized speech and unusual facial expressions. The system should be in a position of tracking these indicators accurately in various forms of data.

We will also do testing of the edge cases to find out how the system behaves in difficult situations, such as noisy frames of video or messed up audio. These conditions can however degrade accuracy, but we are looking for the system to still yield meaningful results. Performance testing will check the speed of the system for processing of large data sets ensuring it can handle the real time or near real time inputs. When the system is trained, we will execute accuracy and assessment tests using such metrics as precision, recall, F1-score,

ROC-AUC to determine the system's efficacy in identifying indicators for schizophrenia. Finally, robustness testing will be employed to ascertain how the system acquires and handles data that may be missing or corrupted for the system to be working properly even at imperfect input.

As we are still in the data processing stage, these tests will finally be implemented after the model is trained. By performing this extensive test we will be able to detect any areas of trouble, optimize performance, and we will be prepared to use the system for real world purposes.

CHAPTER 5

RESULTS AND EVALUATION

5.1 RESULTS

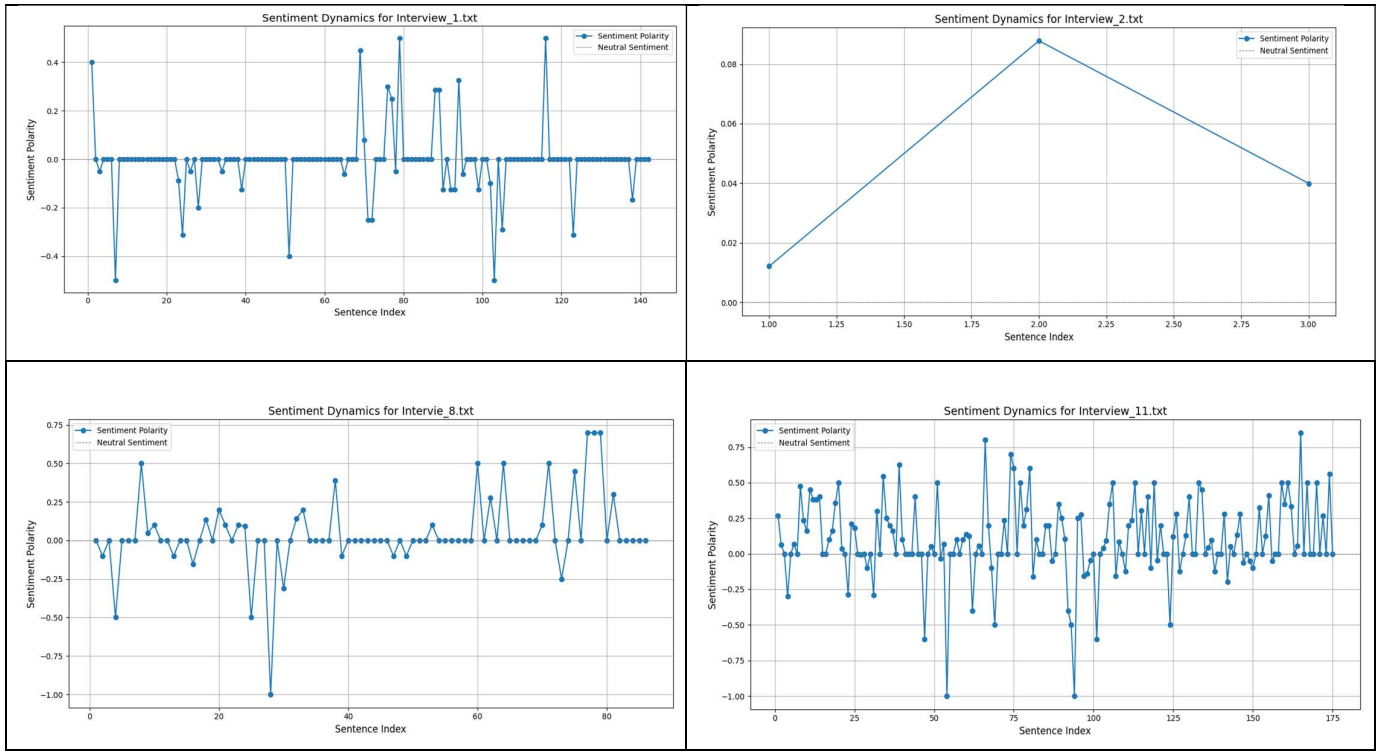
In this chapter we shall describe the expected findings from our work with the multimodal schizophrenia detection system in progress and describe the evaluation process for measuring the performance of the system. At present we are in the data processing stage where we are concentrating on preparing and extracting the relevant features from video, audio, and text data. After the all preprocessing is done, we will now start training and testing the model with varied configurations like varying data size, feature extraction method and epochs of training to evaluate the model's ability to learn efficiently about these markers of schizophrenia.

Because the system is still in development, we have not yet conducted a full evaluation but within the initial tests the focus will be on testing measurable outcomes from system to identify behaviors such as disorganized speech, unusual facial expressions and speech abnormalities. This system will be measured in terms of standard metrics (precision, recall, F1-score, and ROC-AUC) of how well it recognizes schizophrenia-related marker. Moving forward in the next stages, we are set to refine our system with the results and will still be testing to ensure that the system is not only accurate but also robust, eventually ready for practical application within real world clinical settings.


```
{"face_landmarks": [], "pose_landmarks": [{"x": 0.61272948980331 42, "y": 0.19678041338920593, "z": -0.3738836646080017,
"visibility": 0.9999198913574219}, {"x": 0.6191256642341614, "y": 0.17165577411651 61, "z": -0.3488354980945587, "visibility":
0.9998441934585571}, {"x": 0.6247599124908447, "y": 0.17126369476318 36, "z": -0.3489505350589752, "visibility": 0.9998080134391785},
{"x": 0.6302360892295837, "y": 0.17035681009292603, "z": -0.34914615750312805, "visibility": 0.9998210072517395}, {"x":
0.5989447236061096, "y": 0.17173433303833008, "z": -0.3624763488769531, "visibility": 0.9998390674591064}, {"x": 0.5891832709312439,
"y": 0.17146283388137817, "z": -0.3625437319278717, "visibility": 0.999812662601471}, {"x": 0.5807121992111206, "y":
0.1711958944797516, "z": -0.3626003861427307, "visibility": 0.9998512268066406}, {"x": 0.6331467032432556, "y": 0.17538627982139587,
"z": -0.18533578515052795, "visibility": 0.999813973903656}, {"x": 0.5630205273628235, "y": 0.17813867330551147, "z":
-0.2440514862537384, "visibility": 0.9998855590820312}, {"x": 0.6209671497344971, "y": 0.22223308682441711, "z":
-0.30809086561203003, "visibility": 0.9999819993972778}, {"x": 0.5982805490493774, "y": 0.22276289761066437, "z":
-0.32525357604026794, "visibility": 0.9999924898147583}, {"x": 0.6751948595046997, "y": 0.3155308961868286, "z":
-0.025637220591306686, "visibility": 0.9999462366104126}, {"x": 0.5017267465591431, "y": 0.321229487657547, "z":
-0.21104107797145844, "visibility": 0.9999620914459229}, {"x": 0.7153455018997192, "y": 0.5010160207748413, "z":
-0.015018434263765812, "visibility": 0.92516028881073}, {"x": 0.48133209347724915, "y": 0.5444795489311218, "z":
-0.2388213723897934, "visibility": 0.9963904023170471}, {"x": 0.696363091468811, "y": 0.6011176109313965, "z": -0.20888450741767883,
"visibility": 0.7731683850288391}, {"x": 0.6327400803565979, "y": 0.6156842708587646, "z": -0.3152236342430115, "visibility":
0.9510425329208374}, {"x": 0.6939668655395508, "y": 0.6450607180595398, "z": -0.2713247835636139, "visibility": 0.7350907325744629},
{"x": 0.6676729917526245, "y": 0.6416347026824951, "z": -0.3644193112850189, "visibility": 0.9027019739151001}, {"x":
0.6815217137336731, "y": 0.6197056770324707, "z": -0.2906336188316345, "visibility": 0.7407788634300232}, {"x": 0.6744195818901062,
"y": 0.6146414279937744, "z": -0.37236008048057556, "visibility": 0.9056102633476257}, {"x": 0.6756462454795837, "y":
0.6046972870826721, "z": -0.22202645242214203, "visibility": 0.7385371923446655}, {"x": 0.6647485494613647, "y": 0.6055137515068054,
"z": -0.3156178891658783, "visibility": 0.9045904278755188}, {"x": 0.6415541768074036, "y": 0.6203976273536682, "z":
0.08316530287265778, "visibility": 0.9983332753181458}, {"x": 0.5430369973182678, "y": 0.6406663656234741, "z":
-0.08272486180067062, "visibility": 0.9971217513084412}, {"x": 0.7624195218086243, "y": 0.6734777688980103, "z": -0.376785010099411,
"visibility": 0.9214161634445519}, {"x": 0.6669076085090637, "y": 0.7082270383834839, "z": -0.4799082577228546, "visibility":
0.9570348858833313}, {"x": 0.7901772260665894, "y": 0.9772857427597046, "z": -0.27420297265052795, "visibility":
0.7534809112548828}, {"x": 0.6238048076629639, "y": 0.978632926940918, "z": -0.3779154121875763, "visibility": 0.8164417743682861},
{"x": 0.7723867893218994, "y": 0.9879439473152161, "z": -0.2676891088485718, "visibility": 0.7243221402168274}, {"x":
0.5993252396583557, "y": 1.0440120697021484, "z": -0.3652198910713196, "visibility": 0.8217829465866089}, {"x": 0.8402463793754578,
"y": 1.0953199863433838, "z": -0.42123809456825256, "visibility": 0.677043616771698}, {"x": 0.6609095335006714, "y":
1.1305335760116577, "z": -0.512919008731842, "visibility": 0.7524979710578918}]}
```


Fig. 5.1: Face Landmarks and Pose Landmarks of Patient

Table 5.1: Sentimental Dynamics Interviews Table




```
{"chroma_stft": [0.3403692841529846, 0.35439738631248474,
0.35793668031692505, 0.38525933027267456,
0.45531558990478516, 0.5129125714302063,
0.4853273332118988, 0.5102426409721375,
0.4514281451702118, 0.4110354483127594,
0.35076722502708435, 0.32531842589378357], "rmse":
0.03906134143471718, "spectral_centroid":
1920.0218241471869, "spectral_bandwidth":
2135.7919330949676, "rolloff": 3540.569744904285,
"zero_crossing_rate": 0.048350970677111756, "mfcc":
[-359.86627197265625, 125.16435241699219,
6.883918762207031, 26.088729858398438,
14.666792869567871, -2.3907830715179443,
-2.0483837127685547, -5.432600975036621,
-6.765193939208984, -2.587078332901001,
-1.1713333129882812, -4.8054890632629395,
-3.572835683822632]}
```


Fig. 5.2: Features Extracted from Interviewer's Audio

CHAPTER 6

CONCLUSIONS AND FUTURE SCOPE

6.1 CONCLUSION

This project is focused on developing multimodal system for schizophrenia detection, which uses both video, audio and text data in order to provide a comprehensive diagnostic tool. By using modern machine learning approaches, the system is expected to record slight behavioral, acoustic, and linguistic signs of schizophrenia and thus determine it more objectively and on the grounds of a data. The base of the project has been based on heavy data preprocessing and feature extraction to make all three modalities' inputs conducive to integrating them into a single framework.

Presently, we are in the data processing stage during which processing of raw inputs into meaningful features is being done. Facial expressions, gestures, and eye movements are now investigated in the video frames while audio is processed for identification of speech related abnormalities like pitch variation and pause. Text data is being extracted to figure out linguistic patterns as well as coherence. This sequential preparation makes the system ready to integrate insights from multiple modalities well.

After its complete implementation, the system is anticipated to overcome numerous weaknesses of traditional diagnostic means including their subjectivity and their failure to provide standardized criteria. The project has the possibility of helping clinicians make more accurate and earlier diagnoses of schizophrenia by presenting a comprehensive analysis of video, audio and text data. Even, as issues such as data scarcity, synchronization, and computational requirements continue to be encountered, the project setup the preconditions for developing a valid and applicable solution that unites the technology and mental health care.

Adequate mention is given to the fact that in addition to possible clinical benefits, this project thus demonstrates the revolutionary impact of artificial intelligence in addressing the complex issues in the field of mental health. Through the combination of state-of-art and machine learning technologies with multimodal data, the system overcomes the limitations of traditional diagnostic approaches to culminate in a more holistic picture of schizophrenia

related behaviours. The utilization of video, audio, and text enables the system to seize a spectrum of indicators facilitating the improvement of diagnostic reliability and high reliability. Innovatively, this also presents avenues of early intervention which may enhance outcome and minimize the long-term effects on patients and their families. Finally, this project prepares the ground for driving further developments in mental health diagnostics, showing the potential of AI for transformation of healthcare in general and quality of living in particular.

6.2 FUTURE SCOPE

The future of this project has potential for development of schizophrenia detection and mental diagnostics. One of the vital efforts will be to scale the system for larger and varied datasets so that it will generalize well among different population. This scalability will boost reliability and applicability to the real-world scenario of the application. In addition, by adding new data modalities like EEG or fMRI, the diagnostic accuracy of the system can be dramatically improved. By merging these neurological data types, it is possible for the system to generate more insightful analysis of brain patterns relating to schizophrenia thus enhancing its potency.

The next important move will be to roll out the system in clinical situations for the practice of day-to-day care. Integration with a telemedicine platform will increase difficulty of mental health diagnostics accessibility, particularly for patients living in remote or underserved territory. In order to promote its use, explainability features will be added to enable clinicians to know the reasoning of the model's predictions, thereby promoting trust in the model's use. The versatility of the system therefore also creates opportunities to adapt the system to detect other mental health conditions (i.e. depression or bipolar disorder) and thus widens its impact as a holistic mental health tool.

Increased accessibility and usability will also be important. Promoting optimization of the model for lightweight devices (smart phones or tablets) may make the model useful in resource constrained settings. Such optimizations would enable the mental health professional to do the draft assessment immediately, malfunction or not, into real time and hence system would be dynamic and practical in clinical practice. This project could completely change the way of diagnosing of the mental health, even by focusing on the scalability, integration, and usability. It can help close the gap between technology and

psychiatry, bringing advanced tools for schizophrenia detection to the service of wide masses of people across diverse clinical settings.

REFERENCES

[1] G. Premananth, Y. M. Siriwardena, P. Resnik, S. Bansal, D. L. Kelly, and C. Espy-Wilson, "*A Multimodal Framework for the Assessment of the Schizophrenia Spectrum*," in Interspeech, 2024.

[2] H. Göker, "*1D-convolutional neural network approach and feature extraction methods for automatic detection of schizophrenia*," SpringerLink, 2023.

[3] S. Siuly, Y. Guo, and O. F. Alcin, "*Exploring deep residual network based features for automatic schizophrenia detection from EEG,*" SpringerLink, 2023.

[4] C. Y. Chuang, Y. T. Lin, C. C. Liu, L. E. Lee, H. Y. Chang, A. S. Liu, S. H. Hung, and L. C. Fu, "*Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic and Visual Cues,*" IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 987-998, 2023.

[5] A. Kanyal, S. Kandula, V. Calhoun, and D. H. Ye, "*Multi-modal deep learning from imaging genomic data for schizophrenia classification,*" in IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023.

[6] G. S. Chilla, L. Y. Yeow, and Q. H. Chew, "*Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods,*" Scientific Reports, vol.12, pp. 1103-1114, 2022.

[7] A. Das et al., "*An explainable AI approach for schizophrenia detection using multimodal fusion,*" IEEE Access, vol. 10, pp. 78777-78787, 2022.

[8] P. Srivastava et al., "*Multimodal fMRI and EEG deep learning analysis for schizophrenia diagnosis,*"Neurocomputing, vol. 476, pp. 73-84, 2021.

[9] C. R. Phang, F. Noman, H. Hussain, C. M. Ting, and H. Ombao, "*A multi-domain connectome CNN for identifying schizophrenia from EEG patterns,*" IEEE Journal of Biomedical and Health Informatics, vol.24, no. 6, pp. 3109-3119, 2020.

[10] P. T. Krishnan, A. N. J. Raj, P. Balasubramanian, and Y. Chen, "*Schizophrenia detection using Multivariate Empirical Mode Decomposition and entropy measures from multichannel EEG signal,*" Biocybernetics and Biomedical Engineering, vol.

40, pp. 743-754, 2020.

[11] D. Ahmedt-Aristizabal et al., "*Identification of children at risk of schizophrenia via deep learning on EEG responses,*" IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2125-2135, 2020.

[12] R. Desai, P. Porob, and P. Rebelo, "*EEG Data Classification for Mental State Analysis Using Wavelet Packet Transform and Gaussian Process Classifier,*" Wireless Personal Communications, vol. 112, no. 2, pp. 1157-1174, 2020.

[13] S. K. Tikka et al., "*AI-based classification of schizophrenia using EEG and SVM,*" Indian Journal of Psychiatry, vol. 62, no. 6, pp. 704-710, 2020.

[14] N. Ji, L. Ma, H. Dong, and X. Zhang, "*EEG feature extraction using DWT and EMD combined with approximate entropy,*" Brain Sciences, vol. 9, no. 8, pp. 216-225, 2019.

[15] S. Liang et al., "*Classification of First-Episode Schizophrenia Using Multimodal Brain Features,*"Schizophrenia Bulletin, vol. 45, no. 4, pp. 903-914, 2019.

[16] R. Salvador, "*Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia,*"Frontiers in Neuroscience, vol. 13, no. 456, 2019.

[17] M. L. Birnbaum et al., "*A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals,*" Journal of Medical Internet Research, vol. 19, no. 2, pp. e84, 2017.

[18] E. Olejarczyk and W. Jernajczyk, "*Graph-based analysis of brain connectivity in schizophrenia,*" PLoSONE, vol. 12, no. 11, e0188629, 2017.

[19] A. Varshney, C. Prakash, N. Mittal, and P. Singh, "*A Multimodal Approach for Schizophrenia Diagnosis using fMRI and sMRI Dataset,*" in Intelligent Systems Technologies and Applications, 2016.

[20] Y. Yin et al., "*Analyzing the EEG Energy of Quasi Brain Death using MEMD,*" in APSIPA ASC, 2011.

APPENDIX

main_merged.pdf

ORIGINALITY REPORT

7%	6%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	nlistsp.inflibnet.ac.in Internet Source	<1%
3	www.mdpi.com Internet Source	<1%
4	www.medrxiv.org Internet Source	<1%
5	"Pervasive Knowledge and Collective Intelligence on Web and Social Media", Springer Science and Business Media LLC, 2024 Publication	<1%
6	Haiman Guo, Shuyi Jian, Yubin Zhou, Xiaoyi Chen et al. "Discriminative analysis of schizophrenia patients using an integrated model combining 3D CNN with 2D CNN: A multimodal MR image and connectomics analysis", Brain Research Bulletin, 2023 Publication	<1%
7	www.nature.com Internet Source	<1%
8	core-cms.prod.aop.cambridge.org Internet Source	<1%
9	www.isca-archive.org Internet Source	<1%
	assets.researchsquare.com	