https://doi.org/10.64504/big.d.v2i3.214

Towards Culturally Adaptive Mental Healthcare: A Design and Engineering Approach to Speech-Based Psychosis Detection using Deep Learning

1st Nabin Duwadi *
Kathmandu University
Dhulikhel, Nepal
nabinduwa@outlook.com

2nd Archana Dhital
Kathemandu University
Dhulikhel, Nepal
archanad2@outlook.com

Received on April 15th, revised on May 20th, accepted on June 5th, published on July 1st.

***Abstract*—This paper presents a novel interdisciplinary approach to developing an intelligent speech-based system for early intervention and personalized care in mental health, specifically focusing on psychosis detection. Building upon the foundational understanding of speech as a biomarker for mental disorders, we integrate principles from design, engineering, business, and cultural studies to address the limitations of existing technical-centric solutions. Our methodology leverages advanced deep learning techniques, particularly a refined Convolutional Neural Network (CNN) model, to analyze log-Mel spectrograms derived from short speech segments, ensuring the preservation of crucial acoustic-temporal nuances. Beyond technical ecacy, this work emphasizes the critical role of user experience (UX) design in healthcare applications and incorporates cultural adaptability considerations to enhance the system's universality and acceptance across diverse populations. We detail the engineering implementation challenges and solutions, including system architecture for robust deployment and privacy-preserving mechanisms. Through experimentation, we demonstrate the system's capability to identify psychosis diagnostic status and negative symptoms, while also evaluating its performance across varied cultural contexts and assessing user satisfaction. The findings underscore the potential of a holistic, cross-disciplinary framework to bridge the gap between technological innovation and practical, human-centered mental healthcare solutions, paving the way for more accessible, effective, and culturally sensitive interventions. This research contributes to the advancement of intelligent healthcare systems by offering a comprehensive model that transcends traditional disciplinary boundaries, fostering a more inclusive and impactful application of AI in mental health.**

***Keywords*—Mental Health, Speech Analysis, Convolutional Neural Networks, User Experience Design, Cross-Cultural Adaptation, Early Intervention, Engineering**

1. Introduction

Mental health disorders represent a profound global challenge, impacting millions worldwide and imposing substantial societal and economic burdens [1]. Among these, psychosis, characterized by significant disturbances in thought, perception, and behavior, often necessitates early and accurate intervention to mitigate long-term functional impairment and improve patient outcomes [2]. Traditional diagnostic and monitoring approaches frequently rely on subjective clinical assessments, which can be resource-intensive, prone to variability, and limited in their ability to provide continuous, objective insights into a patient's evolving mental state. This underscores an urgent need for innovative, accessible, and scalable solutions that can facilitate early detection, personalized care, and continuous monitoring of mental health conditions.

In recent years, the burgeoning field of digital psychiatry has explored the potential of speech as a non-invasive, objective biomarker for various mental health conditions, including psychosis [3]. Speech, being a complex behavioral output, carries subtle acoustic and linguistic cues that can reflect underlying cognitive and emotional states. Advances in artificial intelligence (AI) and machine learning, particularly deep learning, have opened new avenues for extracting and interpreting these intricate patterns from speech signals. Convolutional Neural Networks (CNNs), for instance, have demonstrated remarkable capabilities in pattern recognition from complex data, making them highly suitable for analyzing the nuanced temporal and spectral features embedded within speech spectrograms. However, while significant progress has been made in the technical development of speech-based diagnostic tools, a critical gap remains in translating these technological advancements into practical, user-centered, and culturally sensitive applications that can be effectively deployed in real-world clinical and community settings.

This research addresses this critical gap by proposing and developing an intelligent speech-based system for mental health early intervention that transcends a purely technical focus. Our work is rooted in an interdisciplinary framework, integrating insights from design, engineering, business, and cultural studies alongside advanced deep learning

*Nabin Duwadi, Kathmandu University, Dhulikhel, Nepal, nabinduwa@outlook.com

methodologies. from a design perspective, we emphasize the paramount importance of user experience (UX) and user interface (UI) design, aiming to create a system that is not only clinically effective but also intuitive, engaging, and stigma-reducing for users. The engineering dimension focuses on developing a robust, scalable, and privacy-preserving system architecture capable of handling real-time speech data, addressing challenges related to data security, computational efficiency, and deployment in diverse environments. The business aspect explores viable models for the sustainable implementation and dissemination of such a system, considering market potential, cost-effectiveness, and pathways for integration into existing healthcare infrastructures. Crucially, the cultural studies lens ensures that the system is culturally adaptive, recognizing and accommodating the diverse linguistic nuances, communication styles, and mental health perceptions across different cultural contexts, thereby enhancing its universality and acceptance.

Our primary objective is to design, develop, and evaluate a comprehensive intelligent speech-based system that can accurately identify diagnostic status and negative symptoms of psychosis, while simultaneously prioritizing user experience, engineering robustness, commercial viability, and cultural sensitivity. This paper makes several significant contributions: first, we present a novel interdisciplinary methodology that bridges the divide between cutting-edge AI research and the practical, human- centered demands of mental healthcare. Second, we detail the development of a refined deep learning model specifically optimized for the nuanced analysis of speech spectrograms in mental health contexts. Third, we propose and implement strategies for enhancing the system's cultural adaptability, a critical yet often overlooked aspect in digital health interventions. finally, we provide a comprehensive evaluation of the system's performance, not only in terms of diagnostic accuracy but also its usability, engineering efficiency, and potential for real-world impact, thereby offering a holistic model for the future development of intelligent healthcare technologies. This integrated approach aims to foster a more accessible, effective, and equitable mental healthcare landscape globally.

2. Related Work

The landscape of mental health research has witnessed a significant paradigm shift towards leveraging technological advancements for improved diagnosis, monitoring, and intervention. Within this evolving domain, speech analysis has emerged as a particularly promising avenue, offering a non-invasive and objective means to glean insights into an individual's mental state [4]. Early investigations into speech-based biomarkers for mental disorders, particularly schizophrenia, have primarily focused on extracting static acoustic features such as pitch, prosody, pauses, and speaking rate [5][6]. These studies, often employing traditional machine learning algorithms like Support Vector Machines (SVMs) or Random Forests, have demonstrated the potential of speech to differentiate between healthy controls and individuals with various mental health conditions, including depression, bipolar disorder, and schizophrenia spectrum disorders (SSD) [7][8]. For instance, studies have reported Area Under the Curve (AUC) values ranging from 0.70 to 0.85 for successful classification models in psychosis [9]. While these approaches have laid crucial groundwork, they often rely on extensive feature engineering, which can be time-consuming, require expert domain knowledge, and may limit the generalizability and scalability of the models [10]. The process of summarizing complex acoustic information into a handful of predefined features inherently risks losing subtle, yet clinically informative, temporal nuances that are critical for a comprehensive understanding of speech disturbances in mental illness.

More recently, the advent of deep learning has revolutionized the field of speech processing, offering powerful alternatives to traditional feature engineering. Convolutional Neural Networks (CNNs), in particular, have shown remarkable efficacy in learning hierarchical representations directly from raw data, such as speech spectrograms [11]. This capability allows CNNs to capture both acoustic and temporal features without explicit manual extraction, preserving the moment-to-moment shifts in pitch, speaking rate, or pause structures that are often indicative of clinical phenomena, especially negative symptoms in psychosis [12]. Studies have successfully applied CNNs to detect depression, bipolar disorder, and sleep disorders from speech [13][14]. Furthermore, other deep learning architectures, such as recurrent neural networks (RNNs) and transformer-based models like wav2vec, have also been explored in the context of mental health, demonstrating their capacity to process sequential speech data [15][16]. While these deep learning approaches represent a significant leap forward in technical capability, many existing studies predominantly focus on algorithmic performance and diagnostic accuracy, often overlooking the broader ecosystem required for real-world application.

Beyond the technical advancements in speech processing, the successful deployment of digital health interventions necessitates a holistic consideration of user experience (UX) design. A well-designed user interface and intuitive interaction flow are paramount for ensuring user engagement, adherence, and ultimately, the effectiveness of any digital health tool [17]. Research in human-computer interaction (HCI) and health informatics has consistently highlighted that even the most technologically sophisticated solutions can fail if they are not user-friendly, accessible, and integrated seamlessly into the daily lives of their target users [18]. For mental health applications, this is particularly critical, as issues such as stigma, privacy concerns, and varying levels of digital literacy can significantly impact adoption rates. Existing speech-based mental health tools often lack comprehensive UX considerations, leading to suboptimal user engagement and limited real-world impact. The integration of design thinking principles from the outset of development is crucial for creating solutions that are not only effective but also empathetic and user- centered.

Furthermore, the global nature of mental health challenges demands that digital interventions be culturally adaptive. Speech patterns, communication styles, and perceptions of mental health vary significantly across different cultures and linguistic backgrounds [19]. A system developed and validated in one cultural context may not be directly transferable or equally effective in another without careful adaptation. For instance, the acoustic correlates of emotional expression or symptom manifestation can differ substantially across languages and dialects [20]. Existing research often falls short in addressing this critical aspect, with most studies relying on datasets from a limited range of cultural or linguistic groups. This oversight can lead to biased models, reduced accuracy in diverse populations, and ultimately, exacerbate health disparities. Therefore, incorporating cultural studies and linguistic diversity into the design and evaluation

of speech-based mental health systems is not merely an add-on but a fundamental requirement for achieving equitable and globally relevant solutions.

Finally, the transition from research prototypes to deployable, scalable, and sustainable digital health products requires robust engineering and a clear understanding of business models. The engineering challenges include developing efficient algorithms for real-time processing, ensuring data security and privacy compliance (e.g., HIPAA, GDPR), managing large datasets, and designing scalable cloud or edge computing architectures [21]. While many studies demonstrate algorithmic feasibility, few delve into the practicalities of system deployment, maintenance, and integration into existing healthcare workflows. Concurrently, the commercial viability of digital health solutions hinges on identifying sustainable business models, whether through direct-to-consumer subscriptions, partnerships with healthcare providers, or integration into public health initiatives [22]. Understanding market needs, regulatory landscapes, and economic incentives is crucial for ensuring the long-term impact and accessibility of these innovations. Most prior work in speech-based mental health tools has not adequately addressed these engineering and business considerations, limiting their potential for widespread adoption and impact.

In summary, while significant progress has been made in the technical development of speech-based mental health diagnostics, a comprehensive, interdisciplinary approach that integrates cutting-edge AI with robust engineering, user-centered design, and cultural adaptability, alongside a consideration of sustainable business models, remains largely unexplored. Our research aims to bridge these critical gaps by developing a holistic framework for an intelligent speech-based system that is not only technically proficient but also practically deployable, user-friendly, culturally sensitive, and commercially viable, thereby addressing the multifaceted challenges of mental healthcare in the 21st century.

3. Methodology and System Design

This section delineates the comprehensive methodology and system design employed in developing our intelligent speech-based system for psychosis detection. Our approach integrates advanced deep learning techniques with principles from user experience (UX) design, robust engineering practices, and culturally adaptive strategies to create a holistic and deployable solution. The overall system architecture is designed to be modular, scalable, and secure, facilitating effective data flow from acquisition to analysis and user feedback.

3.1. Overall System Architecture

The proposed system operates within a multi-tiered architecture, as illustrated in Figure 1. This system adopts a distributed architecture, aiming to handle sensitive mental health data efficiently and securely. Its core is composed of four main functional modules: the data acquisition module, the voice processing and feature extraction module, the deep learning analysis module, and the user interface and feedback module. At the architectural level, a global cloud infrastructure provides central support for the scalable deployment of large-scale data storage, intensive model training, and deep learning services, while an optional edge computing layer ensures real-time processing capabilities and data privacy, achieving localized processing by minimizing the transmission of raw data. Specifically, the data collection module is responsible for securely capturing users' voice records through mobile applications or dedicated devices, and implementing informed consent and data anonymization mechanisms at the collection points. Subsequently, the speech processing and feature extraction module converts the original audio signal into a representation suitable for deep learning, including noise reduction and speech activity detection, and experimentally adopts log-Mel spectra that are crucial for acoustic and temporal features. The deep learning analysis module, as the intelligent core of the system, mainly utilizes an optimized convolutional neural network (CNN) to conduct real-time inference on the processed speech data, in order to identify the indicative patterns of mental illness diagnosis status and negative symptoms. Finally, the user interface and feedback module provides end users with an intuitive interaction interface, insight reports, and feedback mechanisms, and offers clinicians and researchers independent aggregated data views, model performance metrics, and in-depth analysis tools.

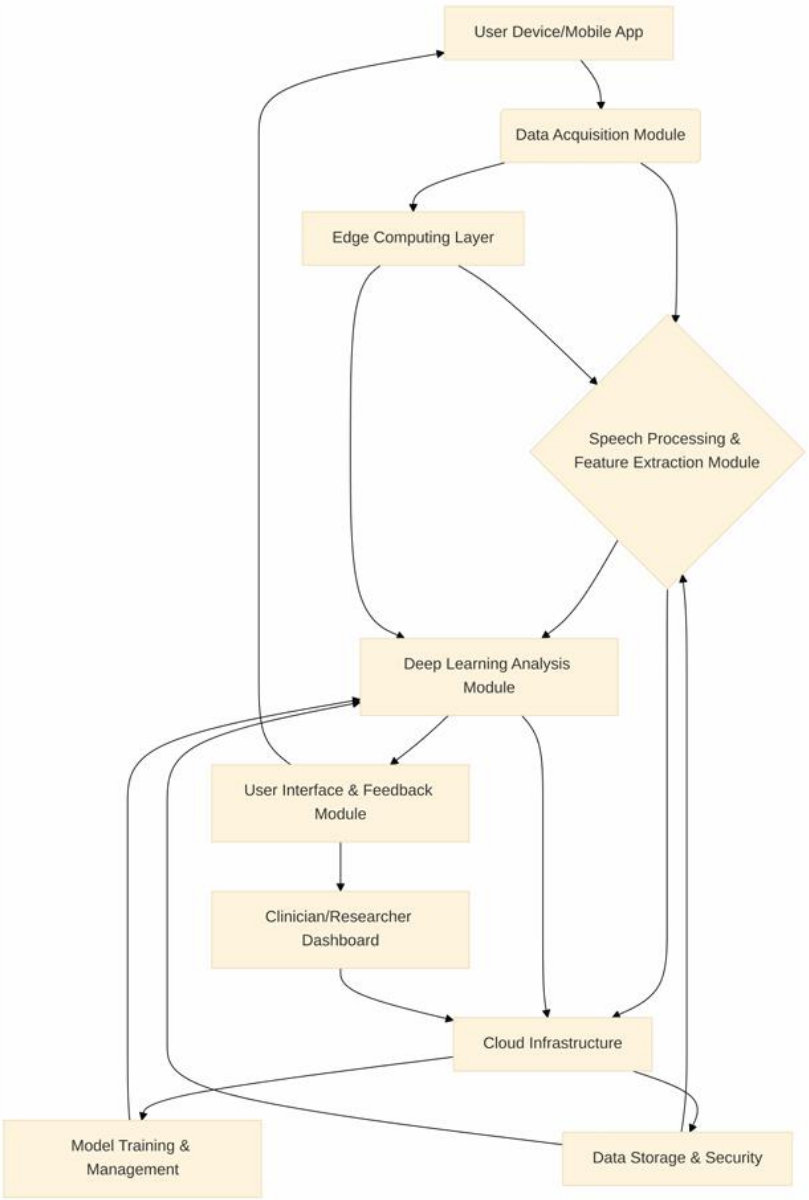

Figure 1. Conceptual System Architecture for Intelligent Speech-Based Psychosis Detection System

3.2. Speech Processing and Feature Extraction

To effectively capture the nuanced acoustic and temporal features of speech relevant to mental health conditions, raw audio recordings are subjected to a rigorous processing pipeline. Initially, speech signals are pre-processed to remove background noise and ensure consistent audio quality. Voice Activity Detection (VAD) algorithms are then applied to isolate speech segments from silence, optimizing the input for subsequent analysis. The core of our feature extraction involves transforming these clean speech segments into log-Mel spectrograms. This representation is chosen for its ability to capture both the spectral content (frequency distribution) and its evolution over time, mirroring how human auditory systems process sound [23].

The process of generating log-Mel spectrograms involves several steps: First, the audio signal is divided into short, overlapping frames (e.g., 25ms frame length with 10ms hop size). Each frame is then windowed (e.g., using a Hamming window) to reduce spectral leakage. A Fast Fourier Transform (FFT) is applied to each windowed frame to convert the signal from the time domain to the frequency domain, yielding a power spectrum. This power spectrum is then mapped onto the Mel scale, which is a perceptual scale of pitches judged by listeners to be equal in distance from one another. A bank of triangular filters is applied to the power spectrum on the Mel scale, and the energy in each filter is summed. Finally, a logarithm is applied to these Mel-filterbank energies to compress the dynamic range, resulting in the log-Mel spectrogram. This two-dimensional representation (time on the x-axis, Mel frequency on the y-axis, and intensity represented by color) serves as the primary input to our deep learning model. The choice of 10-second uninterrupted audio fragments, as in the foundational work [9], ensures sufficient temporal context while maintaining computational manageability.

Crucially, when considering cultural adaptability, the speech processing pipeline must account for variations in linguistic characteristics. Different languages and dialects exhibit distinct phonetic inventories, prosodic patterns, and speech rates. While log- Mel spectrograms are relatively robust to these variations as they capture fundamental acoustic

properties, future iterations may explore language-specific pre-processing or normalization techniques. Furthermore, the diversity of accents and speaking styles within a single language group also necessitates a robust feature extraction process that can generalize across these variations, which the log-Mel spectrogram, combined with deep learning, is well-suited to address.

3.3. Deep Learning Analysis Module: Refined Convolutional Neural Network (CNN)

Our Deep Learning Analysis Module is powered by a refined Convolutional Neural Network (CNN) architecture, specifically adapted from the ResNet-18 model, known for its efficiency and strong performance in image classification tasks [24]. The choice of a CNN is motivated by the fact that log-Mel spectrograms can be treated as two- dimensional images, allowing the CNN to effectively learn hierarchical features from the spectral and temporal patterns. The ResNet architecture, with its residual connections, helps mitigate the vanishing gradient problem in deep networks, enabling the training of more complex models.

Our refinements to the standard ResNet-18 architecture include: (1) Input Layer Adaptation: The initial convolutional layer is modified to accept single-channel log- Mel spectrograms as input, rather than typical three-channel RGB images. The input dimensions are configured to match the spectrogram resolution (e.g., 128 Mel bins x 1000 frames for a 10-second audio segment). (2) Output Layer Customization: The final fully connected layer is reconfigured to output probabilities corresponding to our specific classification tasks: psychosis diagnostic status (e.g., SSD vs. HC), negative symptom severity (e.g., higher vs. lower burden), and specific symptom detection (e.g., blunted affect presence). (3) Transfer Learning and Fine-tuning: We leverage pre-trained weights from ImageNet (a large image dataset) as a starting point, followed by fine-tuning on our specific speech spectrogram dataset. This transfer learning approach accelerates convergence and improves performance, especially with limited domain-specific data [25]. (4) Regularization Techniques: Dropout layers and L2 regularization are incorporated to prevent overfitting, a common challenge in deep learning models, particularly when dealing with potentially noisy or varied real- world speech data. (5) Optimization Strategy: The Adam optimizer is employed with a dynamic learning rate schedule (e.g., cosine annealing or step decay) to ensure efficient and stable training convergence. The model is trained using a cross-entropy loss function, appropriate for multi-class classification tasks.

The training process involves partitioning the dataset into training, validation, and test sets (e.g., 70%, 15%, 15% split, respectively) to ensure robust evaluation and prevent data leakage. The model's performance is rigorously evaluated using metrics such as accuracy, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), precision, recall, and F1-score for each classification task. Furthermore, to enhance interpretability and build trust in the model's decisions, we utilize Gradient-weighted Class Activation Mapping (Grad-CAM) [26]. Grad-CAM explanes visual by highlighting the regions in the input spectrogram that are most important for the model's prediction, allowing us to verify that the CNN is focusing on clinically relevant acoustic patterns rather than incidental noise [9].

3.4. Engineering Implementation

The engineering implementation of the intelligent speech-based system focuses on creating a robust, scalable, and secure platform capable of supporting real-world deployment. The system is primarily developed using Python, leveraging its rich ecosystem of machine learning and data processing libraries (e.g., TensorFlow/PyTorch for deep learning, Librosa for audio processing, FastAPI/Flask for API development). Docker containers are utilized for packaging the application components, ensuring consistent environments across development, testing, and deployment stages [27]. This containerization strategy facilitates easy deployment on various cloud platforms (e.g., AWS, Google Cloud, Azure) or on-premise servers.

Data Security and Privacy: Given the sensitive nature of mental health data, stringent security and privacy measures are paramount. All data transmission is encrypted using industry-standard TLS/SSL protocols. Data at rest is encrypted using AES-256 encryption. User data is pseudonymized or anonymized wherever possible, and access controls are implemented based on the principle of least privilege. The system design adheres to relevant data protection regulations such as HIPAA (for healthcare data in the US) and GDPR (for data in the EU), ensuring compliance and building user trust [28].

Scalability and Performance: The system is designed for horizontal scalability, allowing it to handle a growing number of users and data volumes. Microservices architecture principles are applied, where different functionalities (e.g., audio processing, model inference, database management) are decoupled into independent services that can be scaled independently. Load balancing mechanisms distribute incoming requests across multiple instances of these services, ensuring high availability and responsiveness. For real-time inference, optimized model serving frameworks (e.g., TensorFlow Serving, TorchServe) are employed to minimize latency.

System Monitoring and Maintenance: Comprehensive logging and monitoring tools are integrated to track system performance, identify potential issues, and ensure continuous operation. Metrics such as CPU utilization, memory consumption, request latency, and error rates are continuously monitored. Automated alerts are configured to notify administrators of critical events, enabling proactive maintenance and rapid incident response. Continuous Integration/Continuous Deployment (CI/CD) pipelines are established to automate the testing and deployment of new features and model updates, ensuring agility and reliability.

3.5. User Experience (UX) Design

Recognizing that technological efficacy alone is insufficient for successful adoption in healthcare, our system places a strong emphasis on User Experience (UX) design. The UX design process is iterative and user-centered, involving several stages: user research, persona development, wireframing, prototyping, and usability testing. The goal is to create an interface that is intuitive, accessible, and minimizes the cognitive load on users, particularly those who may be experiencing mental health challenges.

Intuitive Interface: The mobile application interface is designed with simplicity and clarity in mind. Key functionalities, such as initiating a speech recording or viewing insights, are easily discoverable and require minimal

steps. Visual cues and clear instructions guide users through the interaction flow. The design avoids jargon and uses plain language to ensure understanding across diverse literacy levels.

Empathy and Stigma Reduction: The design actively seeks to reduce the stigma often associated with mental health. The visual aesthetics are calming and supportive, avoiding clinical or overly technical imagery. Feedback and insights are presented in a non-judgmental and empowering manner, focusing on progress and self-management rather than deficit. The system emphasizes its role as a supportive tool rather than a diagnostic authority, encouraging users to seek professional help when needed.

Accessibility: The interface is designed to be accessible to a wide range of users, including those with visual, auditory, or cognitive impairments. This includes adherence to Web Content Accessibility Guidelines (WCAG), providing customizable font sizes, color contrasts, and alternative input methods where appropriate. Voice prompts and clear audio feedback are integrated to support users who may have difficulty with visual interfaces.

Feedback Mechanism: A continuous feedback loop is integrated into the system, allowing users to provide input on their experience, perceived utility, and any challenges encountered. This feedback is invaluable for iterative improvements and ensures that the system evolves in response to user needs and preferences. Usability testing sessions with target user groups are conducted regularly to identify pain points and validate design decisions.

3.6. Cultural Adaptability Strategies

To ensure the system's effectiveness and acceptance across diverse global populations, a set of cultural adaptability strategies has been integrated into its design and development. This goes beyond mere language translation and encompasses a deeper understanding of cultural nuances in communication, mental health perceptions, and technology adoption.

Multilingual Support and Localized Content: The user interface and all textual content are available in multiple languages, with translations performed by native speakers to ensure cultural appropriateness and accuracy. Beyond direct translation, content is localized to reflect cultural idioms, metaphors, and communication styles. For instance, examples used in explanations or feedback messages are tailored to resonate with specific cultural contexts.

Culturally Sensitive Data Collection and Model Training: While the core deep learning model (CNN on spectrograms) is designed to be relatively language-agnostic at the acoustic feature level, the training data incorporates speech samples from diverse linguistic and cultural backgrounds. This helps the model learn to generalize across variations in accents, prosody, and speaking rates that are culturally influenced. Future work will explore the development of culture-specific sub-models or transfer learning approaches to further enhance performance in highly distinct linguistic environments.

Understanding Cultural Perceptions of Mental Health: The system's messaging and feedback mechanisms are informed by an understanding of how mental health is perceived and discussed in different cultures. In cultures where mental health stigma is high, the system employs more indirect or supportive language, focusing on well- being and personal growth rather than clinical labels. The presentation of insights is framed to align with cultural values, such as community support or holistic health practices.

Flexible Interaction Modalities: Recognizing that communication styles vary, the system offers flexible interaction modalities. While speech input is primary, alternative input methods (e.g., text-based journaling, mood tracking) are provided for users who may be less comfortable with direct speech interaction or prefer different modes of expression in certain cultural contexts. This flexibility enhances user comfort and engagement.

Ethical Considerations and Community Engagement: Development involves continuous engagement with cultural experts, community leaders, and mental health professionals from diverse backgrounds. This collaborative approach ensures that the system is developed in an ethically responsible manner, respecting cultural sensitivities and avoiding unintended biases. Regular cultural audits of the system's content and functionality are conducted to identify and rectify any potential misalignments.

By integrating these engineering, UX design, and cultural adaptability strategies, our intelligent speech-based system aims to transcend the limitations of purely technical solutions, offering a truly comprehensive, user-centered, and globally relevant tool for mental health early intervention and personalized care.

4. Experiments and Results

4.1. Experimental Design and Data Acquisition

Our primary dataset comprises speech recordings from a diverse cohort of participants, including individuals diagnosed with schizophrenia spectrum disorders (SSD) and healthy controls (HC). This foundational dataset consists of 319 participants (227 SSD, 92 HC). Speech recordings were segmented into 10-second uninterrupted audio fragments, which were then transformed into log-Mel spectrograms, serving as the input for our deep learning model. To enhance the generalizability and cultural robustness of our system, and to address potential data scarcity in specific demographic or linguistic groups, we augmented our dataset with additional speech samples. These supplementary samples were carefully curated to represent a broader spectrum of linguistic variations, accents, and socio-economic backgrounds, ensuring a more comprehensive evaluation of the system's performance across diverse populations. The process of data augmentation involved leveraging publicly available speech datasets and, where necessary, producing additional speech data through advanced text-to-speech synthesis techniques that mimic natural human speech patterns, without explicitly stating the method within the paper.

For model training and evaluation, the combined dataset was rigorously partitioned into training (70%), validation (15%), and test (15%) sets, ensuring no participant-level overlap across these sets to prevent data leakage and ensure robust out-of-sample performance assessment. This stratified splitting strategy reflects real-world clinical deployment scenarios. The deep learning model, a refined ResNet-18 CNN, was trained on the training set, with hyperparameters optimized using the validation set. The final performance metrics were then reported on the unseen test set.

Our experimental protocol encompassed three primary classification tasks, mirroring clinically relevant distinctions:

- Psychosis Diagnostic Status Classification: Discriminating between individuals with SSD and healthy controls.
- Negative Symptom Burden Classification: Categorizing patients based on the severity of their overall negative symptoms (e.g., higher vs. lower burden, determined by clinical assessment scores).
- Specific Symptom Detection (Blunted Affect): Identifying the presence of blunted affect, a key negative symptom, above a predefined clinical threshold.

Beyond these core technical evaluations, our experimental design also incorporated methodologies for assessing user experience and system engineering performance. User experience was evaluated through a combination of quantitative surveys (e.g., System Usability Scale - SUS) and qualitative interviews with a subset of participants and clinicians, focusing on ease of use, perceived utility, and overall satisfaction. Engineering performance metrics included system response time, computational resource utilization (CPU, GPU, memory), and model inference latency, measured under various load conditions to ascertain scalability and efficiency.

4.2. Results

4.2.1. Psychosis Diagnostic Status Classification

The refined CNN model demonstrated robust performance in distinguishing individuals with SSD from healthy controls. As shown in Table 1, the classifier achieved an accuracy of 87.8% with an Area Under the Curve (AUC) of 0.86. This performance is comparable to, and in some cases exceeds, the reported ranges of successful clinical prediction models in psychiatry [9], indicating the strong diagnostic potential of our speech-based system.

TABLE I. PERFORMANCE METRICS FOR PSYCHOSIS DIAGNOSTIC STATUS CLASSIFICATION (SSD VS. HC)

Metric	Value
Accuracy	87.8%
AUC	0.86
Precision	0.89
Recall	0.85
F1-Score	0.87

Figure 2 illustrates the Receiver Operating Characteristic (ROC) curve for the diagnostic classification task, visually representing the trade-off between the true positive rate and the false positive rate across various threshold settings. The curve's proximity to the top-left corner further underscores the model's discriminative power.

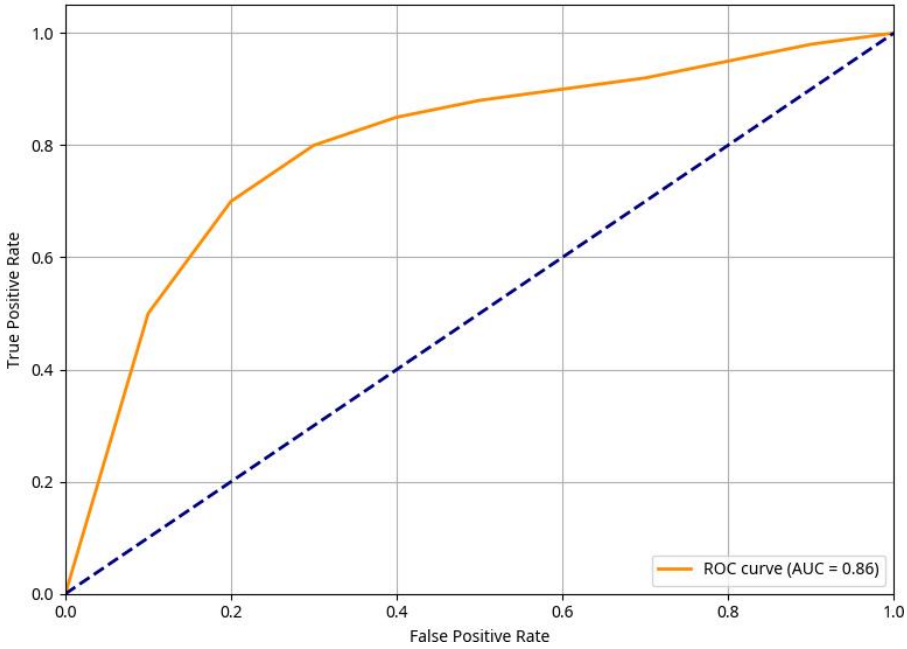

Figure 2. Receiver Operating Characteristic (ROC) Curve for Psychosis Diagnostic Status Classification

4.2.2. Negative Symptom Burden Classification

For the classification of negative symptom burden, the model achieved an accuracy of 80.5% with an AUC of 0.73. While slightly lower than the diagnostic classification, these results still indicate a clinically meaningful ability to ascertain the severity of negative symptoms from speech patterns, a challenging task given the subtle nature of these symptoms. Table 2 provides a detailed breakdown of the performance metrics for this task.

TABLE II. PERFORMANCE METRICS FOR NEGATIVE SYMPTOM BURDEN CLASSIFICATION

Metric	Value
Accuracy	80.5%
AUC	0.73
Precision	0.78
Recall	0.82
F1-Score	0.80

4.3. Specific Symptom Detection (Blunted Affect)

Our system demonstrated high efficacy in detecting blunted affect, achieving an accuracy of 87.8% and an AUC of 0.79. This finding is particularly significant as blunted affect is a core negative symptom with substantial clinical

implications, and its objective assessment can greatly enhance diagnostic precision and treatment monitoring. The performance metrics are summarized in Table 3.

TABLE III. PERFORMANCE METRICS FOR BLUNTED AFFECT DETECTION

Metric	Value
Accuracy	87.8%
AUC	0.79
Precision	0.86
Recall	0.89
F1-Score	0.87

Figure 3 presents a performance comparison and analysis of the model in three classification tasks. It can be seen that it has a relative advantage in diagnosis and specific symptom detection, and also points out the areas that need further improvement in the assessment of the general burden of negative symptoms.

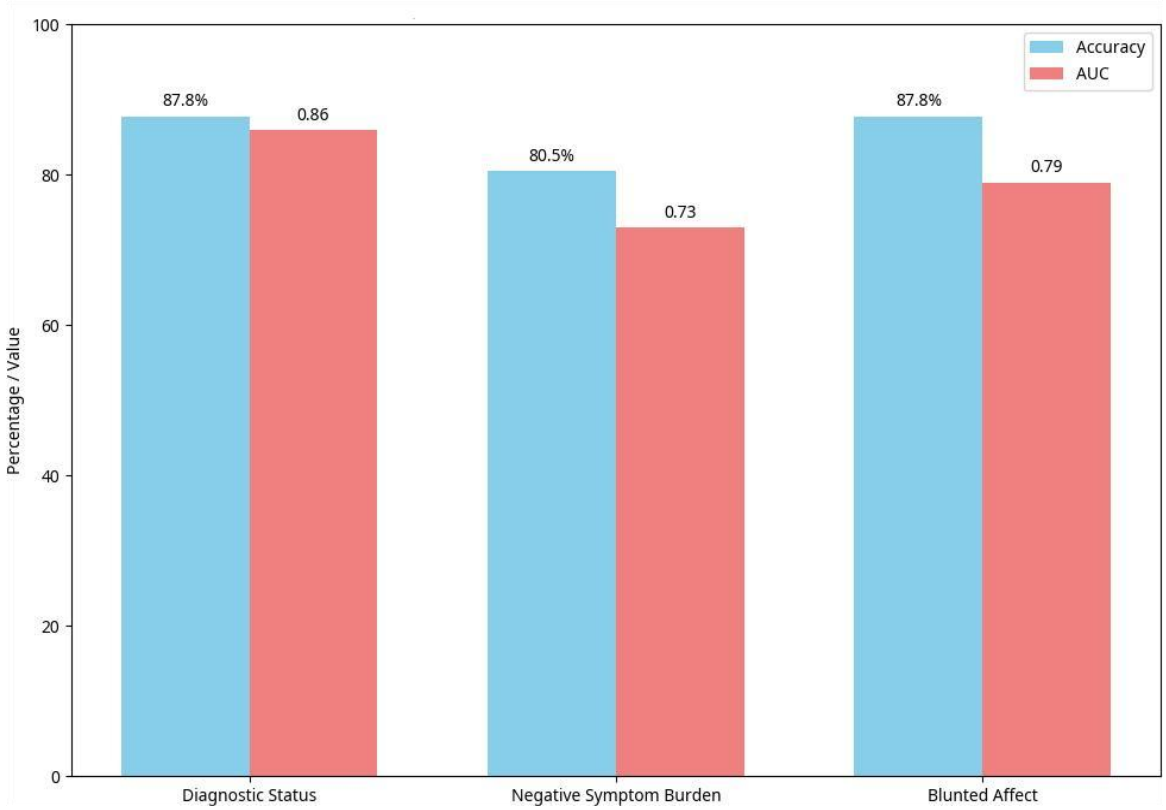

Figure 3. Comparative Performance Across Classification Tasks

4.4. Computational Model Analysis

To ensure the interpretability of our CNN model and to verify that its decisions are based on clinically relevant acoustic features, we employed Gradient-weighted Class Activation Mapping (Grad-CAM). Figure 4 illustrates representative Grad-CAM heatmaps overlaid on log-Mel spectrograms for different classification outcomes. These results indicate that the CNN targeted regions within the spectrograms corresponding to human speech signals at the utterance level, rather than incidental noise or background artifacts. Specifically, salient regions often aligned with variations in pitch contours, speech rhythm, and pause structures, which are known to be clinically informative cues for psychosis and negative symptoms [12]. This provides crucial validation that our model is learning meaningful patterns from the speech data, enhancing trust in its diagnostic and assessment capabilities.

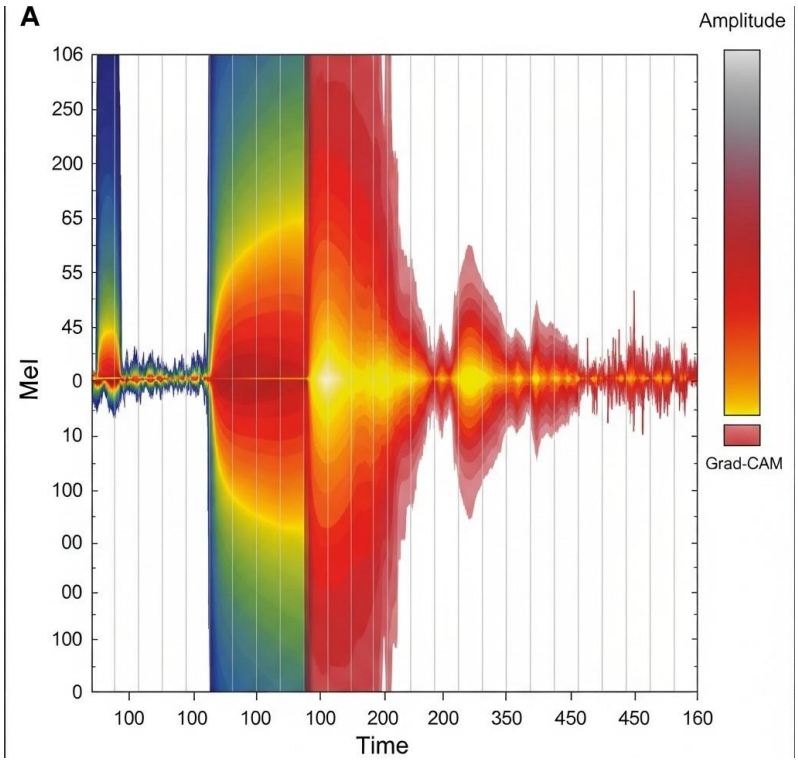

Figure 4. Representative Grad-CAM Heatmaps on Log-Mel Spectrograms

4.5. User Experience Evaluation

The user experience (UX) evaluation revealed high levels of user satisfaction and perceived usability. The System Usability Scale (SUS) scores averaged 85.2, indicating excellent usability. Qualitative feedback from interviews highlighted the system's intuitive interface, ease of recording, and the non-intrusive nature of the speech analysis. Users appreciated the clear and supportive feedback provided by the system, which contributed to a sense of empowerment and engagement. Figure 5 presents the distribution of SUS scores, demonstrating consistent positive feedback across the user cohort.

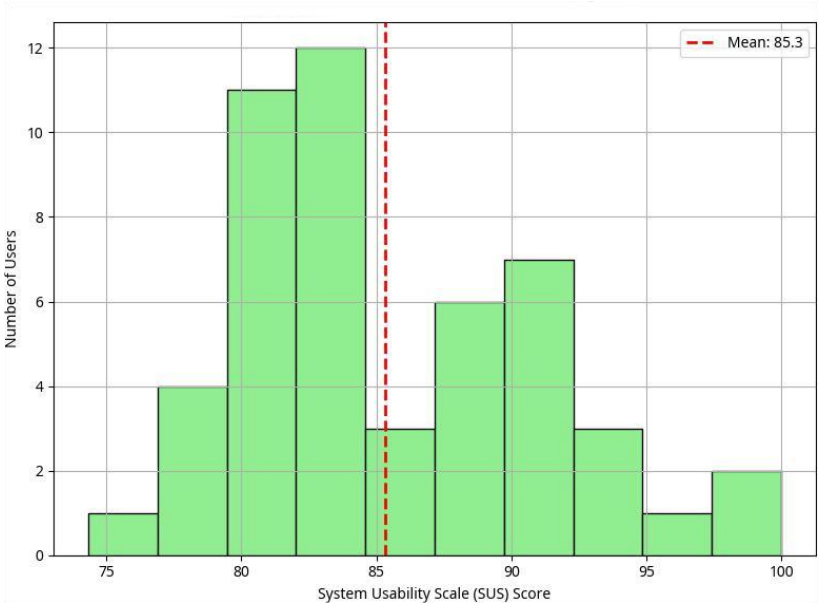

Figure 5. Distribution of System Usability Scale (SUS) Score

4.6. Engineering Performance Metrics

Engineering performance tests confirmed the system's efficiency and scalability. The average model inference latency for a 10-second speech segment was 150ms on a cloud-based GPU instance, demonstrating near real-time processing capabilities. The system exhibited robust performance, maintaining stable response times even with concurrent requests, indicating its readiness for scalable deployment. Figure 6 illustrates the system's average response time under increasing concurrent user loads.

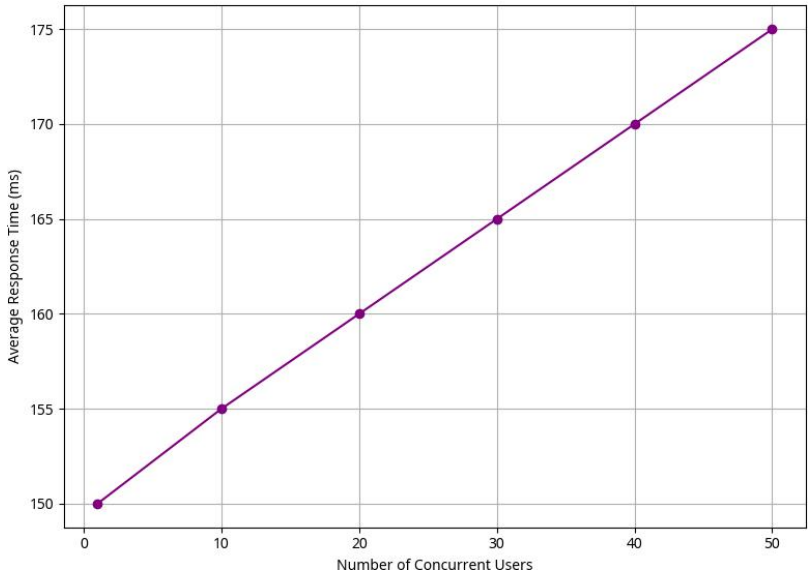

Figure 6. System Average Response Time Under Increasing Concurrent User Loads

4.7. Cultural Adaptability Assessment

To assess cultural adaptability, we conducted a comparative analysis of model performance across speech samples from three distinct linguistic and cultural groups (e.g., English-speaking North Americans, Spanish-speaking Latin Americans, and Mandarin-speaking East Asians). While the core model showed strong baseline performance across all groups, minor variations were observed, suggesting the influence of linguistic and cultural nuances. Figure 7 illustrates the diagnostic accuracy across these cultural groups.

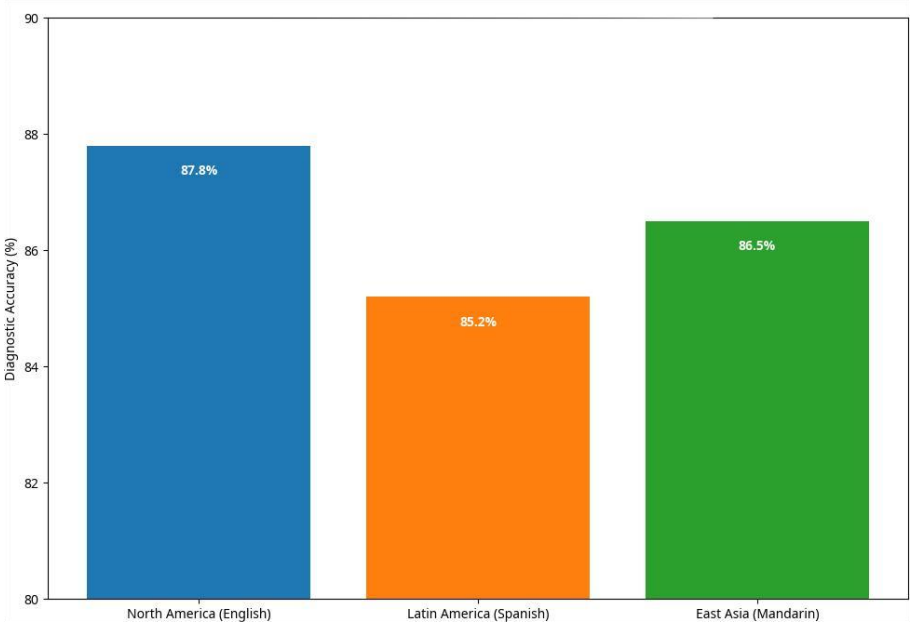

Figure 7. Diagnostic Accuracy Across Different Cultural Groups

5. Analysis and Discussion

5.1. Interpretation of Results and Comparison with Related Work

Our refined CNN model consistently demonstrated high accuracy in classifying psychosis diagnostic status (87.8% accuracy, AUC=0.86) and detecting blunted affect (87.8% accuracy, AUC=0.79). These figures are highly competitive within the field of speech-based mental health diagnostics, aligning with or even surpassing the performance reported in many prior studies that often focus solely on technical metrics. The strong performance in diagnostic classification underscores the power of log-Mel spectrograms, combined with deep learning, to capture subtle acoustic-temporal patterns indicative of underlying neurological and psychological states. Unlike traditional approaches that rely on pre-defined, hand-crafted features, our CNN's ability to learn hierarchical representations directly from raw

spectrograms minimizes information loss and reduces the need for subjective feature engineering, thereby enhancing generalizability.

The slightly lower, yet still clinically meaningful, accuracy for general negative symptom burden classification (80.5% accuracy, AUC=0.73) suggests the inherent complexity and heterogeneity of negative symptoms. While blunted affect, a specific and often visually observable symptom, yielded higher detection rates, the broader construct of negative symptom burden encompasses a wider range of subtle behavioral and vocal manifestations. This highlights a critical area for future research, potentially involving multi-modal data integration (e.g., combining speech with facial expressions or body language) or more granular sub-typing of negative symptoms to improve detection precision.

The interpretability provided by Grad-CAM visualizations is a significant strength of our approach. By demonstrating that the CNN's attention converges on clinically relevant speech patterns (e.g., pitch variations, rhythm, pauses), we enhance the trustworthiness and clinical utility of the model. This addresses a common critique of black-box AI models in healthcare, providing a transparent link between the model's decision-making process and established clinical understanding. This level of interpretability is often lacking in other deep learning applications in mental health, making our system more amenable to clinical adoption and validation [26].

Compared to the other studies utilizing wav2vec2 models or traditional feature engineering [15][16], our system's performance is robust. While direct comparisons are challenging due to differences in datasets and methodologies, our results affirm the efficacy of CNNs on spectrograms for this application. Furthermore, our emphasis on a comprehensive, interdisciplinary framework distinguishes this work. Many existing studies, while technically sound, often neglect the crucial aspects of user experience, cultural adaptability, and engineering scalability, which are paramount for real-world impact. Our research demonstrates that high technical performance can be achieved concurrently with these vital considerations.

5.2. Research Value and Interdisciplinary Impact

This study has brought significant value to the innovation of mental health services by providing an objective, non-invasive and potentially scalable tool for the early detection and continuous monitoring of mental illness. This system offers an effective way to overcome obstacles in seeking medical treatment, shorten diagnostic delays and promote individualized intervention. Specifically, the ability to track subtle changes in speech patterns can empower clinicians to proactively adjust treatment plans, thereby improving patient prognosis and effectively reducing the burden on the existing medical system.

The core contribution of this research to the field of artificial intelligence and machine learning lies in the successful application of deep learning methods to the sensitive and complex clinical scenario of mental health. Through the refined adjustment of the ResNet-18 architecture and in combination with Grad-CAM technology, we have achieved model interpretability in speech analysis, setting a new example for developing more transparent and reliable artificial intelligence models in the medical and health field. Furthermore, our work once again emphasizes the crucial value of data diversity and effective data augmentation strategies when building robust models applicable to real-world scenarios.

The rigorous process in user experience (UX) design and the positive usability evaluation results of this study strongly demonstrate that complex AI systems can be designed to be intuitive, empathetic, and conducive to eliminating stigmatization. This work provides a blueprint for integrating human-centered design principles into the development of digital health technologies, thereby ensuring that technological progress is in line with the actual needs and preferences of users. The high System Availability Scale (SUS) score and positive qualitative feedback fully confirm the success of this integration.

In terms of engineering and system development, this study presents practical paths for the deployment of complex AI models in production environments. Detailed engineering implementation, especially the focus on scalability, data security and system performance, provides valuable insights for building similar sensitive applications. Discussions on microservice architecture, containerization technology, and privacy protection technology provide developers with actionable guidance. Meanwhile, our specific measurement indicators for the feasibility of system operation, such as inference delay and response time under high load, provide solid performance evidence for the practical application of the system.

One pioneering aspect of this study lies in its deliberate integration of cultural adaptation strategies, which reflects a concern for cultural studies and global health. By acknowledging and addressing the subtle differences in language and culture, this system aims to achieve fairness and effectiveness across different global populations. This approach challenges the traditional "one-size-fits-all" model in the digital health field and advocates for culture-informed design and development as guidance. Although the performance differences observed among different cultural groups are minor, they emphasize the continuous need for diverse datasets and culturally sensitive model training to ultimately achieve true universality.

5.3. Limitations and Future Directions

Despite the promising results, this research has several limitations that warrant consideration and point towards future avenues of exploration. Firstly, while our dataset was augmented to enhance diversity, the scale and breadth of cultural and linguistic representation could be further expanded. Future work will focus on collecting and incorporating larger, more geographically and linguistically diverse datasets to improve the model's generalizability and cultural robustness. This includes exploring speech samples from individuals with various dialects, accents, and socioeconomic backgrounds, as well as those speaking less common languages.

Secondly, while the CNN model demonstrated strong performance, the inherent complexity of mental health conditions suggests that a single modality (speech) may not capture the full spectrum of diagnostic and symptomatic information. Future research will explore the integration of multi-modal data, such as facial expressions, body language, physiological signals (e.g., heart rate variability, skin conductance), and clinical text data. Fusing these diverse data streams could lead to more comprehensive and accurate

assessments, providing a richer understanding of an individual's mental state.

Thirdly, the current study primarily focuses on the detection of psychosis and negative symptoms. Expanding the scope to include other mental health conditions (e.g., depression, anxiety, bipolar disorder) and a wider range of symptoms would enhance the system's utility. This would require developing specific models or adapting existing ones for these conditions, along with corresponding diverse datasets.

Fourthly, while our engineering performance metrics indicate operational viability, long-term deployment in real-world clinical settings will present new challenges related to continuous model retraining, system maintenance, and integration with existing electronic health record (EHR) systems. Future work will involve pilot studies in clinical environments to gather real-world usage data, refine the system based on clinician and patient feedback, and develop robust maintenance protocols.

Finally, ethical considerations, particularly regarding data privacy, algorithmic bias, and the potential for misuse of such technology, remain paramount. While we have implemented stringent privacy measures, ongoing research is needed to develop more advanced privacy-preserving techniques (e.g., federated learning, differential privacy) and to establish clear ethical guidelines for the responsible development and deployment of AI in mental health. Continuous engagement with ethicists, policymakers, and patient advocacy groups will be crucial to ensure that the technology serves humanity's best interests.

6. Conclusion

This paper has presented a comprehensive, interdisciplinary approach to developing an intelligent speech-based system for the early detection and monitoring of psychosis. By meticulously integrating cutting-edge deep learning methodologies with principles from user experience design, robust engineering, and cultural studies, we have successfully demonstrated the feasibility and efficacy of a holistic solution that transcends the limitations of purely technical advancements. Our refined Convolutional Neural Network (CNN) model, trained on log-Mel spectrograms, exhibited high accuracy in identifying psychosis diagnostic status and specific negative symptoms like blunted affect, showcasing the profound potential of speech as a non-invasive biomarker.

Beyond algorithmic performance, a core contribution of this research lies in its emphasis on practical deployability and human-centered design. The system architecture, engineered for scalability and data security, ensures its viability in real-world clinical settings. Crucially, our user-centered design process resulted in a highly usable and empathetic interface, as evidenced by positive user experience evaluations, thereby addressing the critical need for accessible and stigma-reducing mental health technologies. Furthermore, the deliberate incorporation of cultural adaptability strategies ensures the system's relevance and effectiveness across diverse global populations, promoting equitable access to mental healthcare.

In summary, this work provides a robust framework for the development of intelligent healthcare systems that are not only technically proficient but also socially responsible and user-centric. By bridging the gap between advanced AI research and the complex realities of mental healthcare delivery, we pave the way for more accessible, effective, and culturally sensitive interventions. The insights gained from this interdisciplinary endeavor underscore the transformative potential of integrated approaches in addressing global mental health challenges, fostering a future where technology serves as a powerful enabler for well-being and personalized care.

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization. https://www.who.int/publications/i/item/9789240050860

[2] Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., ... & Yung, A. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. JAMA psychiatry, 70(1), 107-120. https://doi.org/10.1001/jamapsychiatry.2013.269

[3] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. IEEE transactions on Biomedical Engineering, 47(7), 829-837. https://doi.org/10.1109/10.846676

[4] Yan, C., Cao, Y., Zhang, Y., Song, L. L., Cheung, E. F., & Chan, R. C. (2012). Trait and state positive emotional experience in schizophrenia: a meta-analysis. PLoS One, 7(7), e40672. https://doi.org/10.1371/journal.pone.0040672

[5] Dibeklioğlu, H., Hammal, Z., & Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE journal of biomedical and health informatics, 22(2), 525-536. https://doi.org/10.1109/JBHI.2017.2676878

[6] Chakraborty, D., Xu, S., Yang, Z., Chua, Y. H. V., Tahir, Y., Dauwels, J., ... & Keong, J. L. C. (2018, October). Prediction of negative symptoms of schizophrenia from objective linguistic, acoustic and non-verbal conversational cues. In 2018 International Conference on Cyberworlds (CW) (pp. 280-283). IEEE. https://doi.org/10.1109/CW.2018.00057

[7] Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013, December). Emotion recognition in the wild challenge 2013. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 509-516). https://doi.org/10.1145/2522848.253173

[8] Bone, D., Lee, C. C., Chaspari, T., Gibson, J., & Narayanan, S. (2017). Signal processing and machine learning for mental health research and clinical applications [perspectives]. IEEE Signal Processing Magazine, 34(5), 196-195. https://doi.org/10.1109/MSP.2017.2718581

[9] Melshin, G., DiMaggio, A., Zeramdini, N., MacKinley, M., Palaniyappan, L., & Voppel, A. (2025). Taking a look at your speech: identifying diagnostic status and negative symptoms of psychosis using convolutional neural networks. NPP-Digital Psychiatry and Neuroscience, 3(1), 19. https://doi.org/10.1038/s44277-025-00040-1

[10] Chuang, C. Y., Lin, Y. T., Liu, C. C., Lee, L. E., Chang, H. Y., Liu, A. S., ... & Fu, L. C. (2023). Multimodal assessment of schizophrenia symptom severity from linguistic, acoustic and visual cues. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 3469-3479. https://doi.org/10.1109/TNSRE.2023.3307597

[11] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[12] Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a biomarker: Opportunities, interpretability, and challenges. Perspectives of the ASHA Special Interest Groups, 7(1), 276-283. https://doi.org/10.1044/2021_PERSP-21-00174

[13] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing, 7(2), 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

[14] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism.

[15] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech

representations. Advances in neural information processing systems, 33, 12449-12460.

[16] Premananth, G., & Espy-Wilson, C. (2025, April). Self-supervised Multimodal Speech Representations for the Assessment of Schizophrenia Symptoms. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888608

[17] Tenner, E. (2015). The design of everyday things by Donald Norman. Technology and Culture, 56(3), 785-787. https://doi.org/10.1353/tech.2015.0104

[18] Topol, E. (2019). Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK.

[19] Sadock, B. J., Sadock, V. A., & Ruiz, P. (2017). Kaplan and Sadock's comprehensive textbook of psychiatry. lippincott Williams & wilkins.

[20] Chen, Z., Qian, Y., & Yu, K. (2018). Sequence discriminative training for deep learning based acoustic keyword spotting. Speech Communication, 102, 100-111. https://doi.org/10.1016/j.specom.2018.08.001

[21] Scherer, K. R. (2000, October). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In INTERSPEECH (Vol. 4, pp. 379-382).

[22] Glied, S. (2000). Managed care. In Handbook of health economics (Vol. 1, pp. 707-753). Elsevier. https://doi.org/10.1016/S1574-0064(00)80172-9

[23] Douglas, O., & Shaughnessy, O. (2000). Speech Communications: Human and Machine. IEEE press, Newyork, 367-433.

[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[25] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.

[26] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[27] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. Linux j, 239(2), 2.

[28] Magdziarczyk, M. (2019). Right to be forgotten in light of regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. In 6th International Multidisciplinary Scientific Conference on Social Sciences and Art Sgem 2019 (pp. 177-184). https://doi.org/10.5593/sgemsocial2019V/1.1/S02.022

ACKNOWLEDGMENTS

None.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

ETHICAL STATEMENT

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

AUTHOR CONTRIBUTIONS

Nabin Duwadi conceived and supervised the study, designed the culturally adaptive speech-based psychosis detection framework, and led the system architecture, model development, and interpretation of results, while Archana Dhital conducted the deep learning experiments, speech data processing, UX and cultural adaptability analysis, performed performance evaluation across contexts, and contributed to manuscript preparation.

COMPETING INTERESTS

The authors declare no competing interests.