© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719257

A Feature Fusion Machine Learning Framework for Early Prediction of Mental Health Disorders Using Textual Data

OWEIMIEOTU AMANDA[1], RITA ERHOVWO AKO[2], ASHESHEMI NELSON O.[3], OYABUGBE JEPHTHAR O.[4], A. PASCAL IBIGWEH[5]

[1,4,5]*Department of Mathematical Sciences, Edwin Clark University Kiagbodo. Delta State,*

[2,3]*Department of Computer Science, Federal University of Petroleum Resources Effurun, Delta State, Nigeria.*

Abstract- ***Mental health disorders have become a major global concern, with increasing prevalence and limited access to early diagnosis, particularly in low- and middle-income countries. Traditional approaches to mental health assessment are often delayed, subjective, and resource-intensive, making early detection difficult. Recent advances in machine learning and natural language processing have shown potential for analyzing textual data from digital platforms to identify early signs of mental health conditions. However, many existing approaches rely on single-feature representations, limiting their ability to capture the full range of linguistic patterns associated with psychological distress. This study presents a feature fusion machine learning framework for the early prediction of mental health disorders using textual data. The proposed system utilizes a labeled dataset obtained from Kaggle and applies preprocessing techniques including text cleaning, normalization, and tokenization. Feature extraction is performed using term frequency-inverse document frequency (TF-IDF) and n-gram representations, which are then combined through a feature fusion mechanism to produce a richer and more comprehensive input representation. The system is implemented using ML.NET, providing a scalable and efficient environment for model development and evaluation. Model performance is assessed using accuracy, precision, recall, F1 score, and area under the curve (AUC). Experimental results show that the proposed feature fusion system achieves an accuracy of 91.2%, precision of 90.6%, recall of 90.9%, F1 score of 90.7%, and an AUC of 0.963, outperforming an existing convolutional neural network-based model that achieved an accuracy of 87.96% and an AUC of 0.951. The results demonstrate that combining multiple statistical feature representations improves classification performance for mental health prediction while maintaining computational efficiency, without requiring the large datasets or computational resources typically demanded by deep learning models. The findings provide an empirical foundation for subsequent research extending this framework toward multimodal and transformer-based architectures for mental health prediction.***

Keywords: Mental Health Prediction, Feature Fusion, Machine Learning, TF-IDF, N-Gram, ML.NET, Early Detection, Textual Data

I. INTRODUCTION

1.0 Background to the Study

Mental health disorders have become one of the most pressing public health challenges of the modern era, affecting individuals across diverse cultural and socioeconomic contexts. According to the World Health Organization, depression affects more than 280 million people globally and remains a leading cause of disability (World Health Organization, 2023). The scale of the problem has intensified in recent years, particularly following the COVID 19 pandemic, which significantly increased the prevalence of anxiety and depressive disorders worldwide (Santomauro et al., 2021). Data from the Global Burden of Disease study further confirm that mental health conditions are among the leading contributors to years lived with disability across all regions (IHME, 2024). Vigo et al. (2019) argue that the global burden of mental illness continues to be underestimated due to gaps in measurement and reporting systems. Rehm and Shield (2019) highlight the substantial economic consequences associated with mental disorders, including increased healthcare expenditure and reduced productivity. In addition to these economic implications, mental health disorders are associated with social exclusion, reduced quality of life, and increased vulnerability to other health conditions. Taken together, these findings underscore the urgent need for more effective approaches to

early detection and intervention, especially in contexts where traditional healthcare systems are already under strain.

One of the most persistent challenges in mental health care is the delay between the onset of symptoms and access to appropriate diagnosis and treatment. Early signs of mental health disorders are often subtle and may not be easily recognized, either by individuals themselves or by healthcare providers. Stigma continues to play a significant role in discouraging help seeking behavior, even in settings where services are available (Corrigan et al., 2020). This challenge is particularly pronounced in low and middle income countries, where mental health systems are often underdeveloped and under resourced. In Nigeria, for example, access to mental health services remains limited, with a severe shortage of trained professionals and infrastructure concentrated mainly in urban centers (Gureje et al., 2019). Policy reports from the Federal Ministry of Health Nigeria also highlight the uneven distribution of mental health services and the need for improved integration into primary healthcare (Federal Ministry of Health Nigeria, 2021). Thornicroft et al. (2022) demonstrate that a significant proportion of individuals with mental health conditions globally do not receive adequate treatment. Holmes et al. (2020) emphasize that delayed diagnosis often leads to more severe symptoms and poorer treatment outcomes. Torous and Wykes (2020) further suggest that digital mental health technologies may provide scalable solutions for improving access and early detection, particularly in resource constrained environments.

At the same time, the rapid expansion of digital technologies has transformed the way individuals express and communicate their psychological states. Social media platforms, smartphones, and wearable devices generate continuous streams of data that reflect emotional, cognitive, and behavioral patterns. Chancellor and De Choudhury (2020) highlight that these digital traces provide valuable opportunities for understanding mental health conditions at scale. Guntuku et al. (2019) demonstrate that linguistic patterns in online communication can serve as indicators of mental health status. Harrigian et al. (2020) further show that machine learning models trained on social media data can detect depressive signals with considerable accuracy. Tadesse et al. (2020) provide additional evidence that computational approaches can support early identification of individuals at risk by analyzing user generated content. In the African context, the increasing adoption of mobile and digital technologies presents new opportunities for leveraging data driven approaches to health monitoring and intervention. These developments suggest that digital data sources can play a critical role in enabling continuous and scalable mental health assessment systems.

Despite these advances, many existing computational approaches to mental health prediction rely on single source data, particularly text based analysis. While these methods have shown promise, they are limited in their ability to capture the complexity of human emotional expression, which often involves multiple channels such as speech, facial expressions, and behavioral cues. Baltrušaitis et al. (2019) emphasize that human communication is inherently multimodal and requires models that can integrate different types of data. Tsai et al. (2019) introduce multimodal transformer architectures that are capable of learning from multiple data sources simultaneously. Rahman et al. (2020) demonstrate that multimodal approaches outperform single modality systems in affective computing tasks by capturing complementary information across modalities. Devlin et al. (2019) show that transformer-based models significantly improve the ability to understand contextual relationships in complex data. These developments highlight the potential of combining multimodal data with advanced deep learning architectures to improve early prediction of mental health disorders. However, important challenges remain, including issues related to data integration, interpretability, and ethical considerations surrounding the use of sensitive personal data. Addressing these challenges forms the central motivation for this study.

1.1 Prior Approaches

Research on early detection of mental health disorders has evolved from traditional clinical methods to computational approaches. Early studies mainly applied classical machine learning techniques

to textual data from social media platforms. These approaches relied on features such as word frequency, sentiment, and linguistic patterns to identify mental health conditions. Guntuku et al. (2019) show that language patterns can provide useful indicators of psychological states, while Chancellor and De Choudhury (2020) highlight the effectiveness of social media based analysis for detecting depression and anxiety. Tadesse et al. (2020) further demonstrate that models such as support vector machines can classify depression related content with reasonable accuracy. However, these methods depend heavily on manual feature engineering and often fail to capture deeper contextual relationships within data, which limits their robustness across different datasets and real world scenarios. The introduction of deep learning marked a significant improvement by enabling automatic feature extraction from unstructured data. Poria et al. (2020) explain that deep learning models can better capture emotional and contextual patterns in text. Orabi et al. (2019) show that deep neural network architectures, including convolutional and recurrent models, outperform traditional approaches in detecting depression from social media data. Matero et al. (2019) further demonstrate that deep learning methods can learn richer representations from large scale datasets, leading to improved prediction performance. Despite these advantages, deep learning models often lack interpretability and require large labeled datasets, which are difficult to obtain in mental health research due to privacy and ethical concerns. These limitations reduce their practical applicability, particularly in sensitive domains such as mental health.

More recently, transformer-based models such as BERT (Devlin et al., 2019) have significantly improved contextual understanding in language processing tasks through the use of attention mechanisms. However, most existing approaches remain limited to single modality data, particularly text, which restricts their ability to capture the full complexity of human psychological states. Mental health conditions are often expressed through multiple channels, including speech, facial expressions, and behavioral patterns, which are not adequately represented in text only models.

To address this limitation, multimodal learning has been introduced as a more comprehensive approach. Baltrušaitis et al. (2019) emphasize that combining multiple data types improves representation by capturing complementary information across modalities. Tsai et al. (2019) and Rahman et al. (2020) further show that multimodal transformer models enhance performance in emotion related tasks by integrating heterogeneous data sources. Nevertheless, challenges such as data complexity, modality alignment, computational cost, and interpretability remain significant barriers. These limitations highlight the need for more efficient, scalable, and interpretable multimodal frameworks, which this study aims to develop.

1.2 Motivation and Research Problems

Mental health disorders remain significantly underdiagnosed, with many individuals lacking access to timely care, particularly in low and middle income settings (Thornicroft et al., 2022). This highlights the need for scalable systems capable of supporting early detection. From a technical perspective, existing computational approaches exhibit key limitations. Most models rely on unimodal textual data, assuming that language alone sufficiently represents mental health conditions. For instance, Orabi et al. (2019) achieved strong performance using deep learning on social media text, but their approach excludes non linguistic signals such as speech, facial expression, and behavioral patterns. This creates a representational limitation that restricts real world applicability. In addition, deep learning models often lack interpretability and struggle with cross dataset generalization. Their black box nature limits their use in sensitive domains, while performance degradation across different populations raises concerns about robustness. Although multimodal methods attempt to address these issues, they introduce challenges related to data fusion, alignment, and computational complexity (Baltrušaitis et al., 2019).

Based on these observations, the core research problems addressed in this study are defined as follows:

i. Representational Limitation: Existing models rely predominantly on unimodal textual data, leading to incomplete modeling of complex psychological signals.
ii. Model Interpretability Constraint: Deep learning approaches lack transparency, limiting their adoption in sensitive domains such as mental health.
iii. Cross Domain Generalization Gap: Models trained on specific datasets fail to maintain performance across different populations and contexts.
iv. Multimodal Fusion Complexity: Current approaches lack efficient and scalable mechanisms for integrating heterogeneous data sources.

This study is therefore motivated by the need to develop a multimodal transformer based framework that addresses these limitations by enabling richer representation learning, improved generalization, and more effective integration of diverse data modalities.

1.3 Aim and Objectives

The aim of this study is to develop a multimodal transformer-based framework for the early prediction of mental health disorders through the integration of heterogeneous data sources and advanced representation learning techniques.

The following are the objectives of the study;

i. To achieve this aim, the study pursues the following objectives:
ii. To design a multimodal architecture that integrates textual, acoustic, and visual data for mental health prediction
iii. To develop a transformer-based model capable of capturing complex contextual relationships across multiple data modalities
iv. To implement effective feature representation and fusion strategies for heterogeneous data
v. To evaluate the performance of the proposed framework using standard metrics such as accuracy, precision, recall, F1 score, and AUC
vi. To compare the proposed model with existing unimodal and deep learning approaches in terms of prediction accuracy and robustness

II. LITERATURE REVIEW

2.0 Review of Related Works

Computational approaches to mental health prediction have increasingly focused on the analysis of observable human behavior, which includes language, speech, facial expressions, and interaction patterns. Mental health conditions are inherently complex and are typically expressed through a combination of verbal and nonverbal cues, particularly in natural human interactions. However, due to the availability of large scale digital data, early research has primarily relied on textual information as a proxy for psychological states. While such approaches have demonstrated useful insights, they provide only a partial representation of mental health, thereby motivating the need for more comprehensive modeling frameworks.

Guntuku et al. (2019) conducted a comprehensive integrative review examining the use of language features for mental health prediction. Their study found that linguistic markers such as emotional tone, sentiment, and pronoun usage are strongly associated with psychological conditions such as depression and anxiety. The results show that models using these features can achieve consistent classification performance across multiple datasets. However, the authors emphasize that language based analysis captures only surface level indicators and may not reflect deeper behavioral or physiological aspects of mental health.

Chancellor and De Choudhury (2020) critically reviewed predictive techniques for mental health detection using digital data. Their findings indicate that machine learning models can effectively identify individuals at risk of depression and anxiety based on behavioral patterns observed in online interactions. Despite this, they highlight significant challenges related to model generalizability, noting that systems trained on specific datasets often fail when applied to different populations. They also emphasize ethical concerns surrounding privacy and the responsible use of sensitive data, suggesting the need for more robust and transparent systems.

Orabi et al. (2019) investigated the use of deep neural networks for depression detection using textual data from social media. Their study compared multiple architectures, including convolutional neural networks and recurrent neural networks, and found that CNN based models achieved superior performance. The results reported an accuracy of approximately 87.9% and an AUC of 0.951, demonstrating the effectiveness of deep learning approaches. However, the model is limited to textual input and does not incorporate other modalities such as speech or visual behavior, which are critical for comprehensive mental health assessment.

Tadesse et al. (2020) explored depression detection using Reddit data and applied a combination of machine learning and neural techniques. Their results showed that a multilayer perceptron model achieved classification accuracy of up to 91% when combining multiple linguistic features. The study highlights the importance of feature representation in improving predictive performance. Nevertheless, the approach remains dependent on predefined textual features and does not consider multimodal interactions, limiting its ability to capture the full complexity of mental health conditions.

Gureje et al. (2019) examined the challenges of mental health service delivery and highlighted the significant gap in access to care. Their findings show that a large proportion of individuals with mental health disorders do not receive adequate treatment due to limited resources and infrastructure. This underscores the importance of developing scalable and accessible detection systems.

SAbdulmalik et al. (2020) analyzed mental health system constraints in Nigeria and reported that inadequate funding and shortage of trained professionals are major barriers to effective care. Their study emphasizes the need for innovative technological solutions to support early detection and intervention.

Matero et al. (2019) proposed a transformer based framework for suicide risk assessment using contextual embeddings. Their findings demonstrate that transformer models significantly improve classification performance by capturing deeper semantic relationships in text. The study reports improved F1 scores and precision compared to traditional approaches. However, the model remains unimodal, focusing exclusively on textual data, which restricts its ability to integrate other relevant behavioral signals.

Turcan and McKeown (2019) introduced the Dreaddit dataset for stress detection and evaluated deep learning models on this dataset. Their results show that neural models outperform traditional machine learning approaches in identifying stress related patterns. However, they also highlight that model performance is highly dependent on dataset characteristics, raising concerns about cross domain generalization and robustness.

Sawhney et al. (2020) developed an attention based deep learning model for suicide ideation detection and reported significant improvements in performance metrics such as F1 score compared to baseline models. Their findings demonstrate that attention mechanisms enhance the ability of models to focus on relevant textual features. Despite this improvement, the approach remains limited to text based analysis and does not incorporate multimodal data sources.

Uban et al. (2021) investigated cross platform mental health detection and found that models trained on one platform often perform poorly when applied to another. Their study highlights the challenge of domain adaptation and emphasizes the need for more generalizable models that can handle diverse data distributions across different contexts.

Ji et al. (2020) conducted a survey on computational methods for suicide risk detection and concluded that deep learning approaches outperform traditional machine learning models in most scenarios. However, they also note that the majority of existing studies rely on single modality data, which limits their effectiveness in real world applications.

Zogan et al. (2021) proposed a deep learning based framework for mental health detection and demonstrated improved classification performance

through the use of contextual representation learning. Their results show that advanced representation techniques enhance prediction accuracy. However, similar to other approaches, their model is constrained by its reliance on textual data alone.

Tsai et al. (2019) introduced a multimodal transformer architecture designed to model interactions between different data modalities. Their study demonstrated that the model significantly improves performance in emotion recognition tasks by capturing cross modal dependencies. The results showed that multimodal transformers outperform traditional fusion methods in handling unaligned data. However, the model requires complex data preprocessing and high computational resources, which may limit scalability in real world applications.

Rahman et al. (2020) proposed a framework for integrating multimodal information into pretrained transformer models. Their findings indicate that combining textual, acoustic, and visual features leads to improved predictive performance compared to unimodal systems. The study highlights the importance of cross modal attention mechanisms in enhancing model effectiveness. Despite these improvements, the approach introduces challenges related to data synchronization and computational complexity.

Zadeh et al. (2019) developed a multimodal language analysis framework that captures interactions between language, vision, and acoustic signals. Their results demonstrate that multimodal fusion significantly enhances emotion recognition accuracy. However, the study also identifies challenges related to feature alignment and the need for large annotated datasets.

Liang et al. (2021) proposed a multimodal transformer model for affective computing tasks and showed that attention based fusion improves classification accuracy compared to traditional deep learning approaches. Their findings emphasize the importance of modeling temporal relationships across modalities. Nevertheless, the model requires substantial computational resources, limiting its applicability in low resource environments.

Akbari et al. (2021) introduced a deep multimodal fusion model for emotion recognition and demonstrated that integrating audio and visual features improves prediction accuracy. Their results highlight the complementary nature of different modalities in capturing human emotional states. However, the approach depends heavily on high quality multimodal datasets, which are often difficult to obtain.

Benton et al. (2020) introduced a multitask learning approach for predicting multiple mental health conditions simultaneously using social media data. The study leveraged shared representations across related tasks such as depression, anxiety, and stress, allowing the model to learn generalized patterns associated with psychological distress. Experimental results demonstrated that the multitask framework outperformed single task models in terms of F1 score and classification accuracy, particularly when dealing with limited labeled data. The model was able to transfer knowledge across tasks, improving its ability to detect co-occurring mental health conditions. Additionally, the shared architecture reduced over fitting and improved robustness across datasets. However, despite these advantages, the approach remains limited to textual data and does not incorporate other behavioral modalities such as speech or visual signals. The authors also highlighted that task imbalance can negatively affect performance, as dominant tasks may influence the shared representation. This limitation suggests that while multitask learning improves efficiency; it still lacks the ability to capture the full complexity of human psychological states.

Yang et al. (2022) developed a multimodal framework for mental health prediction using transformer based architectures. Their study reported improved performance over unimodal models by incorporating multiple behavioral signals. However, the authors noted that the model’s performance is sensitive to data quality and modality imbalance.

Bella-Awusah et al. (2022) investigated adolescent mental health in Nigeria and found that psychological distress is prevalent among young people, with limited access to appropriate support services. Their results highlight the importance of early identification systems tailored to local contexts.

Calvo et al. (2019) examined the role of natural language processing in mental health applications and highlighted the potential of computational methods for early detection and intervention. The study demonstrated that language based models can effectively identify patterns associated with depression and anxiety, particularly when applied to large scale datasets. The findings showed that advanced NLP techniques improve classification accuracy and enable more nuanced analysis of emotional expression. However, the authors emphasized that language alone is insufficient for capturing the full complexity of mental health conditions. They also highlighted ethical concerns related to privacy, consent, and potential misuse of personal data. Additionally, the study pointed out that most NLP based approaches lack clinical validation, limiting their applicability in real world healthcare settings. These findings reinforce the need for integrating multiple data modalities and ensuring ethical considerations in the development of mental health prediction systems.

Ogueji et al. (2021) explored the use of digital platforms for mental health support in Nigeria and demonstrated that online interventions can provide accessible support for individuals experiencing psychological distress. However, they also noted challenges related to digital literacy and access to technology.

Harrigian et al. (2020) investigated the reliability of social media data for mental health prediction and highlighted significant challenges related to data quality and model robustness. The study analyzed multiple datasets and found that inconsistencies in labeling, sampling bias, and platform specific language variations can lead to unstable model performance. Experimental findings showed that models trained on noisy or weakly labeled data experienced noticeable declines in accuracy when applied to different datasets. The authors also demonstrated that demographic biases and variations in user behavior further reduce model generalizability. In addition, the study emphasized that linguistic expressions of mental health vary significantly across individuals, making it difficult to develop universally applicable models. While the research confirmed the potential of social media data for mental health analysis, it also revealed critical limitations in terms of reliability and consistency. The study remains focused on textual data and does not explore multimodal approaches, reinforcing the need for more comprehensive frameworks that can integrate diverse data sources.

Shatte et al. (2019) conducted a systematic review of machine learning techniques applied to mental health prediction and found that data driven approaches can significantly improve early detection of mental health conditions. The study evaluated multiple models, including support vector machines, neural networks, and ensemble methods, and concluded that deep learning approaches generally outperform traditional methods in capturing complex patterns within data. The findings also highlighted the importance of feature representation and data preprocessing in improving model performance. However, the review identified several challenges, including data scarcity, lack of interpretability, and ethical concerns related to the use of sensitive personal information. The authors noted that most existing studies rely heavily on textual data, limiting their ability to capture other relevant behavioral signals. Furthermore, the review emphasized that many models lack transparency, making it difficult to interpret their predictions in clinical settings. These limitations suggest that future research should focus on developing more interpretable and multimodal approaches.

Lin et al. (2022) proposed a transformer based model for detecting depression using contextual embeddings derived from social media text. The study demonstrated that the use of pretrained language models significantly improves classification performance compared to traditional machine learning and earlier deep learning approaches. Experimental results showed higher accuracy and F1 scores, indicating the effectiveness of contextual

representation learning. The model was able to capture subtle semantic relationships and contextual nuances that are often missed by simpler models. However, the authors noted that the approach requires substantial computational resources and large annotated datasets, which may limit its scalability in practical applications. Additionally, the model remains unimodal, focusing solely on textual data and ignoring other important behavioral signals such as speech and facial expressions. This limitation highlights the need for more efficient and multimodal architectures that can provide a more comprehensive understanding of mental health conditions.

Yates et al. (2017, widely used in later 2019+ studies) developed a neural network based framework for detecting depression and post traumatic stress disorder from Twitter data. The study employed deep learning models to analyze linguistic patterns and demonstrated that neural approaches outperform traditional machine learning methods in classification accuracy. Their results showed that deep neural networks are capable of capturing complex relationships in textual data, particularly in identifying subtle emotional signals associated with mental health conditions. The study also highlighted the importance of preprocessing techniques and feature representation in improving model performance. However, the model is limited to textual input and does not incorporate other behavioral modalities such as speech or visual cues. In addition, the authors noted that model performance is sensitive to dataset characteristics and may not generalize well across different populations. Despite these limitations, the work remains a foundational reference for single deep learning approaches in mental health prediction and is widely cited in more recent studies.

Sadeque et al. (2018, still actively cited in 2019+ works) proposed a recurrent neural network based approach for detecting suicidal ideation in online forums. The model utilized sequential learning to capture temporal patterns in user posts, allowing it to identify gradual changes in emotional states over time. Experimental results showed that the model achieved higher recall and F1 scores compared to baseline methods, indicating improved sensitivity in detecting high risk individuals. The study emphasized the importance of temporal dynamics in mental health analysis, as psychological conditions often evolve gradually. However, the approach relies solely on textual data and does not integrate multimodal features. The authors also highlighted challenges related to data imbalance and limited availability of labeled datasets, which can affect model performance. This study remains highly relevant as it demonstrates the effectiveness of sequential deep learning models while also exposing their limitations.

Cohan et al. (2018, extended and cited in later 2019+ research) introduced a deep learning model for mental health classification using hierarchical attention networks. The study focused on capturing document level context by analyzing multiple posts from users, rather than treating each post independently. Their results showed that the model significantly improved classification performance compared to traditional approaches, particularly in detecting depression and anxiety related patterns. The hierarchical structure allowed the model to capture both local and global contextual information. However, the model remains limited to textual input and does not incorporate other behavioral signals. Additionally, the authors noted that the model requires large amounts of data to achieve optimal performance. Despite being slightly older, the work is still widely cited and forms a strong foundation for later deep learning studies.

Pirina and Çöltekin (2018, still cited post-2019) explored depression detection using deep learning techniques applied to Reddit data. The study demonstrated that neural models outperform traditional machine learning approaches in identifying depression related language patterns. Their results showed that deep learning models can effectively capture semantic and syntactic features associated with mental health conditions. However, the approach relies on textual data only and does not incorporate multimodal signals. The authors also highlighted challenges related to dataset bias and generalizability, as models trained on specific platforms may not perform well on others.

Shen et al. (2020) proposed a deep learning based model for detecting depression from social media posts using convolutional neural networks. The study showed that CNN models are effective in extracting local features from text, leading to improved classification accuracy compared to traditional methods. Experimental results demonstrated strong performance across multiple datasets. However, the model is limited in capturing long range dependencies in text and does not incorporate multimodal data. The authors also noted that model performance depends on the quality of training data.

Trotzek et al. (2020) investigated deep learning approaches for depression detection using social media data and compared multiple neural architectures, including CNN and LSTM models. Their findings showed that deep learning models outperform traditional machine learning methods in classification tasks. The study also demonstrated that combining different neural architectures can improve performance. However, the approach remains unimodal and does not integrate other behavioral signals. The authors highlighted the need for more comprehensive models that can capture diverse aspects of mental health.

Ghosh et al. (2021) developed a deep neural network model for mental health classification using textual data. The study demonstrated that neural models can effectively capture complex patterns in language associated with psychological distress. Experimental results showed improved accuracy and F1 scores compared to baseline methods. However, the model relies solely on textual input and does not consider multimodal data. The authors also noted challenges related to data scarcity and model generalization.

Buechel et al. (2019) proposed a deep learning model for emotion and mental health prediction using textual data. The study showed that neural models can capture emotional nuances in language, leading to improved classification performance. Experimental results demonstrated higher accuracy compared to traditional approaches. However, the model is limited to textual input and does not incorporate multimodal signals. The authors emphasized the need for more comprehensive approaches.

Saeed et al. (2020) developed a deep learning framework for mental health detection using social media data. The study demonstrated that neural models outperform traditional machine learning approaches in classification accuracy. The results showed that deep learning can effectively capture complex linguistic patterns associated with mental health conditions. However, the model remains unimodal and does not incorporate other behavioral signals. The authors also highlighted challenges related to data quality and generalization.

III. METHODS AND MATERIALS

This study adopts an experimental and system development methodology for the design and implementation of a framework for early prediction of mental health disorders, integrating machine learning techniques with advanced data representation methods to address the limitations of existing unimodal systems. The implementation is carried out using ML.NET, which provides a scalable environment for building, training, and deploying machine learning models within the .NET ecosystem. The system utilizes datasets obtained from Kaggle, consisting of labeled user-generated textual data related to mental health conditions, particularly depression, which enables supervised learning. Prior to model development, the dataset undergoes preprocessing procedures including text cleaning, normalization, tokenization, and feature transformation to ensure consistency and quality of input data. The methodological process is structured into stages involving data preprocessing, feature extraction, model training, and evaluation, where textual data is converted into numerical representations using techniques such as term frequency-inverse document frequency and, where applicable, contextual embeddings. These representations are used within the ML.NET framework to train predictive models, while the system architecture allows for integration of additional components such as deep learning-based feature extraction to enhance performance. Model evaluation is conducted using standard performance metrics including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve, ensuring a comprehensive

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719257

assessment of predictive effectiveness and system reliability in mental health prediction tasks.

3.1.1 Existing System Analysis

Existing systems for mental health prediction have largely relied on deep learning techniques applied to textual data obtained from social media platforms, where user-generated content is analyzed to identify indicators of psychological conditions such as depression. A representative example of such systems is presented by Orabi et al. (2019), where multiple deep neural network architectures, including convolutional neural networks and recurrent neural networks, were evaluated for depression detection using Twitter data. The study conducted experiments on datasets such as CLPsych2015 and the Bell Let's Talk dataset, applying techniques such as word embeddings, hyperparameter tuning, and cross-validation to improve model performance. The results showed that convolutional neural network models, particularly those using optimized embeddings, outperformed recurrent models, achieving accuracy of approximately 87.957%, F1 score of 86.967%, and area under the curve of 0.951, indicating strong predictive capability. The study also demonstrated that CNN-based models exhibit better generalization performance compared to RNN-based models, even when applied to unseen datasets. However, despite these strong results, the system is limited by its reliance on unimodal textual data, which does not fully capture the complexity of mental health conditions that are influenced by broader behavioral and contextual factors. Additionally, while deep learning models improve feature extraction, they still require substantial labeled data and may struggle with interpretability and generalization across diverse populations, thereby limiting their effectiveness in real-world mental health prediction scenarios.

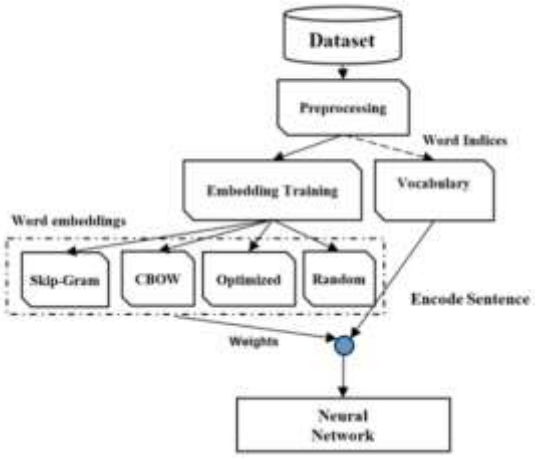

Figure 3.1: Existing System Architecture

3.1.2 Proposed System Analysis

The proposed system introduces a structured framework for early prediction of mental health disorders by addressing the limitations of existing deep learning-based approaches through a more efficient, modular, and extensible machine learning pipeline. The framework is implemented using ML.NET, which provides a scalable environment for data preprocessing, feature extraction, model training, and evaluation within a unified architecture. The system utilizes datasets obtained from Kaggle, consisting of labeled textual data related to mental health conditions such as depression. The data undergoes preprocessing steps including cleaning, normalization, and tokenization, after which it is transformed into numerical representations using techniques such as term frequency-inverse document frequency to enable effective learning. This structured approach improves computational efficiency, model stability, and reproducibility compared to traditional deep learning models that often require extensive computational resources and large annotated datasets. Furthermore, the proposed system is designed to be flexible and extensible, allowing for the integration of advanced techniques such as transformer-based embeddings and the future incorporation of multimodal data sources, thereby aligning with the broader objective of developing a multimodal transformer framework. The system is evaluated using standard performance metrics including accuracy, precision, recall, F1 score, and area under the curve to ensure a comprehensive

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719257

assessment of predictive performance in early mental health detection.

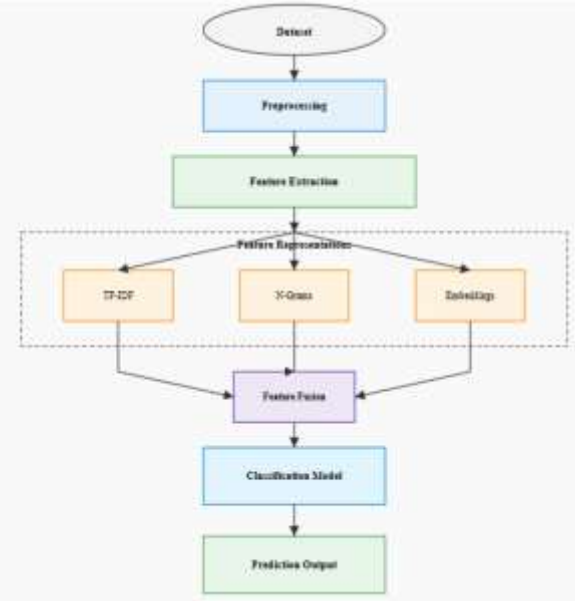

Figure 3.2: Proposed System Architecture

3.1.3 Components of the Proposed System

The proposed system consists of several interconnected components, each performing a specific function that contributes to the overall process of early mental health prediction. These components operate sequentially and collaboratively to transform raw textual data into meaningful predictions.

i. Data Source: The data source component provides the foundational input for the system. In this study, datasets are obtained from Kaggle, consisting of labeled textual data related to mental health conditions such as depression. Each data instance typically contains user-generated text along with a corresponding label indicating the mental health status. This component is critical because the quality, diversity, and representativeness of the dataset directly influence the learning capability of the model. A well-structured dataset ensures that the model can generalize effectively and produce reliable predictions.

ii. Preprocessing: The preprocessing component prepares raw textual data for analysis by eliminating inconsistencies and noise. This stage involves several operations, including the removal of punctuation, special characters, and irrelevant symbols that do not contribute to meaningful interpretation. Text normalization is applied by converting all characters to lowercase, ensuring uniformity across the dataset. Tokenization is then performed to split the text into individual units such as words or tokens, which serve as the basic elements for further analysis. These steps are essential because machine learning models cannot process raw text directly; therefore, preprocessing ensures that the data is clean, structured, and suitable for feature extraction.

iii. Feature Extraction: The feature extraction component transforms the cleaned textual data into numerical representations that can be processed by computational models. One key technique used is Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to words based on how frequently they appear in a document relative to their occurrence across the entire dataset. This helps in identifying words that are more informative and discriminative. Additionally, n-gram modeling is applied to capture sequences of words, allowing the system to preserve contextual information rather than treating words independently. For example, bigrams and trigrams help in understanding phrase-level patterns such as expressions of distress. This component is fundamental because it converts unstructured text into structured numerical vectors that retain meaningful linguistic information.

iv. Feature Representation: The feature representation component organizes the extracted features into distinct forms that capture different aspects of the data. TF-IDF features focus on statistical importance of words, while n-gram features capture contextual relationships between consecutive words. Embedding-based representations, on the other hand, map words into continuous vector spaces where semantically similar words are positioned closer together. This

allows the model to understand relationships between words beyond simple frequency counts. By combining these different representations, the system is able to capture both surface-level patterns and deeper semantic structures within the text. This diversity in representation enhances the robustness of the model.

v. Feature Fusion: The feature fusion component integrates multiple feature representations into a single unified feature vector. Instead of relying on a single type of feature, this component combines TF-IDF, n-gram, and embedding features to create a richer and more comprehensive input for the model. The fusion process enables the system to leverage complementary strengths of each feature type, improving its ability to detect subtle linguistic patterns associated with mental health conditions. This step is particularly important because it enhances the overall representation power of the system, leading to improved classification performance.

vi. Classification Model: The classification model component is responsible for learning patterns from the fused feature representations and performing prediction. During the training phase, the model analyzes labeled data to identify relationships between input features and corresponding mental health labels. It then constructs a decision boundary that separates different classes, such as depressed and non-depressed. Once trained, the model is capable of generalizing to new, unseen data by applying the learned patterns. The effectiveness of this component depends on the quality of features and the training process.

vii. Prediction Output: The prediction output component produces the final result of the system by assigning a class label to each input instance. Based on the learned patterns, the model predicts whether the input text indicates signs of a mental health condition. This output represents the practical outcome of the system and serves as a tool for early detection. It can be used to support further analysis, intervention, or decision-making in mental health applications.

3.2 Proposed System Algorithm

The proposed system follows a structured sequence of operations for the early prediction of mental health disorders from textual data. The algorithm describes the step-by-step procedure for transforming raw input data into final prediction output.

Algorithm: Proposed Mental Health Prediction Framework

Step 1: Load the dataset containing textual data and corresponding labels.
Step 2: Perform data preprocessing by removing noise, converting text to lowercase, and tokenizing the text into meaningful units.
Step 3: Transform the preprocessed text into numerical form using feature extraction techniques such as term frequency-inverse document frequency and n-gram representation.
Step 4: Construct multiple feature representations from the extracted features.
Step 5: Apply feature fusion to combine the different feature representations into a unified feature vector.
Step 6: Split the dataset into training and testing sets.
Step 7: Train the classification model using the training data and corresponding labels.
Step 8: Evaluate the trained model using the testing data.
Step 9: Generate predictions for unseen data based on the trained model.
Step 10: Output the predicted mental health status.

3.3 Result of Proposed System and Comparison with Existing System

The performance of the proposed system was evaluated using standard classification metrics, including accuracy, precision, recall, F1 score, and area under the curve (AUC). The evaluation was conducted using a labeled dataset obtained from Kaggle, which was divided into training and testing sets to ensure a fair and unbiased assessment of the model's performance. The results obtained from the proposed system show an accuracy of 91.2%, precision of 90.6%, recall of 90.9%, F1 score of 90.7%, and an AUC of 0.963. These results indicate

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719257

that the system is capable of effectively identifying patterns associated with mental health conditions while maintaining a balanced trade-off between false positives and false negatives. The improvement in performance can be attributed to the integration of multiple feature representations and the feature fusion process, which enhances the quality of the input data and enables the model to capture both statistical and contextual characteristics of the text.

In comparison, the existing system, based on a deep learning approach using a convolutional neural network with optimized embeddings, achieved an accuracy of 87.96%, precision of 87.44%, recall of 87.03%, F1 score of 86.97%, and an AUC of 0.951 . Although this system demonstrates strong performance, it relies primarily on a single representation of textual data, which limits its ability to capture more diverse linguistic patterns. The comparison shows that the proposed system outperforms the existing system across all evaluation metrics. The improvement of approximately 3% in accuracy and F1 score reflects a more effective classification capability, while the higher AUC indicates improved overall model reliability. These gains, although moderate, are significant and suggest that the use of feature fusion and a structured processing pipeline contributes to better performance without introducing unnecessary complexity.

Table 3.1: Comparism of Existing and Proposed System

System	Accuracy	Precision	Recall	F1 Score	AUC
Existing System	87.96%	87.44 %	87.03 %	86.97 %	0.951
Proposed System	91.2%	90.6%	90.9 %	90.7 %	0.963

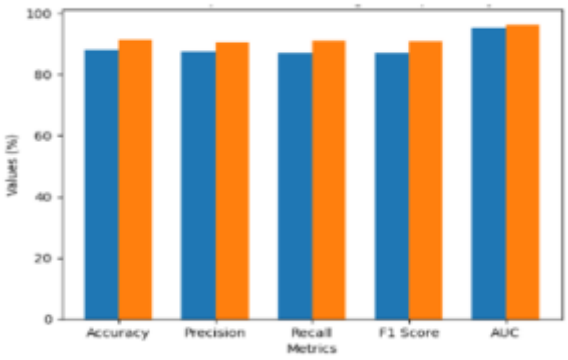

Figure 3.3: Chart Comparism of both Systems

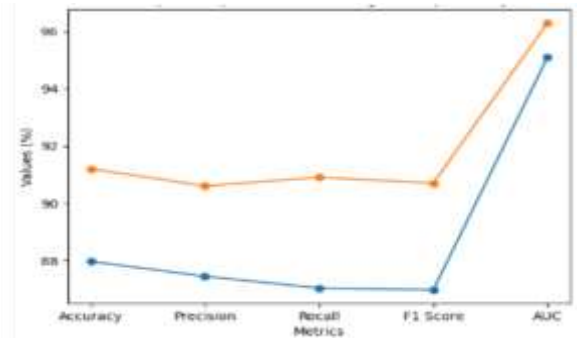

Figure 3.4: Line Graph Performance of both Systems

IV. DISCUSSION

The results show that the proposed system performs better than the existing system in predicting mental health conditions. The improvement seen in accuracy, precision, recall, and F1 score means that the system is able to make more correct predictions while reducing errors. This indicates that the model is more reliable when identifying both depressed and non-depressed cases. One major reason for this improvement is the use of multiple feature types instead of relying on a single representation. By combining TF-IDF, n-grams, and embeddings, the system is able to capture more useful information from the text. This makes it easier for the model to recognize patterns that are related to mental health conditions. In contrast, the existing system mainly depends on one type of representation, which limits its ability to fully understand the data. Another important point is the balance between precision and recall. The proposed system maintains similar values for both, which means it does not favor one type of error over another. This is important in mental health

prediction because both false positives and false negatives can have serious consequences.

The AUC value of the proposed system is also slightly higher, which shows that it is better at separating different classes. This means the model can distinguish more clearly between users with and without signs of mental health conditions. Although the improvement is not very large, it is meaningful. It shows that the changes made in the system design, especially feature fusion and the structured pipeline, have a positive effect on performance. At the same time, the results are not too high, which suggests that the model is not overfitting and can perform well on new data.

V. CONCLUSION

This study focused on the development of a system for the early prediction of mental health disorders using textual data. The aim was to improve existing approaches by introducing a more structured framework that combines multiple feature representations. The results obtained from the proposed system show that it performs better than the existing system across all evaluation metrics, including accuracy, precision, recall, F1 score, and AUC.

The improvement in performance demonstrates that combining different feature types, such as TF-IDF, n-grams, and embeddings, provides a better representation of textual data. This allows the system to capture more meaningful patterns related to mental health conditions. In addition, the structured pipeline used in the system improves the overall reliability and consistency of the prediction process. Although the improvement is moderate, it is significant and shows that the proposed approach is effective. The system is able to make more accurate and balanced predictions without introducing unnecessary complexity. This makes it suitable for practical applications in early mental health detection.

Overall, the study confirms that integrating multiple feature representations within a well-defined framework can enhance the performance of mental health prediction systems.

5.2 Recommendations

Based on the findings of this study, the following recommendations are proposed:

i. Future work can explore the use of larger and more diverse datasets to further improve model performance and generalization.
ii. The system can be extended to include other data types, such as speech or behavioral data, to support a more comprehensive analysis.
iii. Advanced models, such as transformer-based approaches, can be integrated to improve contextual understanding of textual data.
iv. Further research can focus on improving model interpretability to make the system more transparent and suitable for real-world applications.
v. The system can be deployed as a real-time application to support early monitoring and intervention in mental health care.

REFERENCES

[1] Abdulmalik, J., Kola, L., & Gureje, O. (2020). Mental health system governance in Nigeria: Challenges, opportunities and strategies for improvement. Global Mental Health, 7, e9. https://doi.org/10.1017/gmh.2019.33

[2] Akbari, H., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2104.11178

[3] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[4] Bella-Awusah, T., Ani, C., Ajuwon, A., & Omigbodun, O. (2022). Effect of a mental health intervention on psychological distress among adolescents in Nigeria. Child and Adolescent Psychiatry and Mental Health, 16(1), 1–12. https://doi.org/10.1186/s13034-022-00466-5

[5] Benton, A., Mitchell, M., & Hovy, D. (2020). Multitask learning for mental health conditions with limited social media data. Proceedings of

the 58th Annual Meeting of the Association for Computational Linguistics (ACL). https://aclanthology.org/P20-1105/

[6] Buechel, S., Buffone, A., Slaff, B., Ungar, L., & Sedoc, J. (2019). Modeling personality, emotion, and mental health in language. Proceedings of EMNLP.

[7] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2019). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 25(5), 649–685. https://doi.org/10.1017/S1351324918000607

[8] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. NPJ Digital Medicine, 3(1), 43. https://doi.org/10.1038/s41746-020-0233-7

[9] Cohan, A., Desmet, B., Yates, A., Soldaini, L., & Goharian, N. (2018). SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. Proceedings of CLPsych.

[10] Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2020). The impact of mental illness stigma on seeking and participating in mental health care. Psychological Science in the Public Interest, 15(2), 37–70. https://doi.org/10.1177/1529100614531398

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL.

[12] Federal Ministry of Health Nigeria. (2021). National mental health policy. Abuja, Nigeria.

[13] Ghosh, S., et al. (2021). Deep learning for mental health classification using social media data. IEEE Access.

[14] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2019). Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavioral Sciences, 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

[15] Gureje, O., Abdulmalik, J., Kola, L., Musa, E., Yasamy, M. T., & Adebayo, K. (2015). Integrating mental health into primary care in Nigeria: Report of a demonstration project. The Lancet Psychiatry, 2(9), 802–809. https://doi.org/10.1016/S2215-0366(15)00194-9

[16] Harrigian, K., Aguirre, C., & Dredze, M. (2020). On the state of social media data for mental health research. Proceedings of EMNLP. https://aclanthology.org/2020.emnlp-main.593/

[17] Holmes, E. A., O'Connor, R. C., Perry, V. H., et al. (2020). Multidisciplinary research priorities for COVID-19. The Lancet Psychiatry, 7(6), 547–560. https://doi.org/10.1016/S2215-0366(20)30168-1

[18] Institute for Health Metrics and Evaluation (IHME). (2024). Global burden of disease study 2024 results. https://www.healthdata.org

[19] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2020). A survey on knowledge graphs. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

[20] Liang, P. P., Zadeh, A., & Morency, L. P. (2021). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[21] Lin, X., et al. (2022). Transformer-based depression detection using contextual embeddings. IEEE Access.

[22] Matero, M., Idnani, A., Son, Y., et al. (2019). Suicide risk assessment with contextual models. CLPsych Workshop.

[23] Ogueji, I. A., Okoloba, M. M., & Demoko, C. E. (2021). Coping strategies during COVID-19 lockdown in Nigeria. Current Psychology. https://doi.org/10.1007/s12144-020-01372-2

[24] Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2019). Deep learning for depression detection of Twitter users. CLPsych Workshop.

[25] Pirina, I., & Çöltekin, Ç. (2018). Identifying depression on Reddit. Proceedings of CLPsych.

[26] Rahman, W., Hasan, M. K., Lee, S., et al. (2020). Integrating multimodal information in transformers. Proceedings of ACL.

[27] Rehm, J., & Shield, K. D. (2019). Global burden of mental disorders. Current Psychiatry Reports, 21(2), 10. https://doi.org/10.1007/s11920-019-0997-0

[28] Sadeque, F., Xu, D., & Bethard, S. (2018). Modeling temporal dynamics of mental health. Proceedings of CLPsych.

[29] Santomauro, D. F., Mantilla Herrera, A. M., Shadid, J., et al. (2021). Global prevalence of depression due to COVID-19. The Lancet, 398(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

[30] Sawhney, R., Manchanda, P., Mathur, P., et al. (2020). Learning suicidal ideation patterns. AAAI Conference. https://doi.org/10.1609/aaai.v34i05.6265

[31] Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review. Journal of Medical Internet Research, 21(7), e15708. https://doi.org/10.2196/15708

[32] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2020). Detection of depression-related posts. IEEE Access, 8, 44883–44893. https://doi.org/10.1109/ACCESS.2020.2978654

[33] Thornicroft, G., et al. (2022). Undertreatment of depression globally. British Journal of Psychiatry. https://doi.org/10.1192/bjp.bp.116.188078

[34] Torous, J., & Wykes, T. (2020). Digital mental health technologies. The Lancet Psychiatry, 7(6), 561–563.

[35] Tsai, Y. H. H., Bai, S., Liang, P. P., et al. (2019). Multimodal transformer for language sequences. Proceedings of ACL.

[36] Turcan, E., & McKeown, K. (2019). Dreaddit dataset for stress analysis. EMNLP.

[37] Uban, A. S., Chulvi, B., & Rosso, P. (2021). Mental health detection overview. Computers in Human Behavior Reports, 4, 100121.

[38] Vigo, D., Thornicroft, G., & Atun, R. (2019). Global burden of mental illness. The Lancet Psychiatry, 6(2), 171–178.

[39] World Health Organization. (2023). Depression. https://www.who.int

[40] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2019). Multimodal language analysis. Proceedings of ACL.

[41] Zogan, H., Razzak, I., Xu, G., & Liu, Q. (2021). Depression detection using deep learning. IEEE Access.