Detection of Mental Health and Depression using Deep Learning

1st Dr.Parul Madan
Associate Professor, Dept. of computer Science and Engineering
Graphic Era Deemend to Be University
Uttarakhand, India

2nd Ansh Badoni
Student, Computer Science and Engineering
Graphic Era Deemend to be University
Uttarakhand, India

3rd Rajeev Sharma
Student, Computer Science And Engineering
Graphic Era Deemed to Be University
Uttarakhand, India

4th Sarthak Singh Rawat
Student, Computer Science And Engineering
Graphic Era Deemed to Be University
Uttarakhand, India

***Abstract*—Mental health disorders pose a significant challenge worldwide, necessitating early and accurate detection for effective intervention and improved outcomes. Traditional diagnostic methods, however, often fall short in terms of scalability and objectivity. In response to these challenges, this research presents a comprehensive analysis of three advanced deep learning approaches for text-based mental health classification: a TextCNN model integrated with bi-directional LSTM, a model utilizing TextCNN with K-fold cross-validation, and a fine-tuned BERT-based model. The study highlights the incremental improvements in classification accuracy achieved by these methods, with the BERT-based model demonstrating superior performance in identifying and categorizing mental health states such as anxiety, depression, suicidal tendencies, stress, and normal. The research underscores the potential of these models to facilitate real-time monitoring and assessment of mental health, offering a scalable and objective tool to assist mental health professionals in making timely and informed decisions.**

I. Introduction

Mental health disorders represent a significant global health challenge, characterized by persistent alterations in cognition, emotion, behavior, or a combination thereof. These disorders encompass a wide spectrum of conditions, including but not limited to depression, anxiety disorders, bipolar disorder, schizophrenia, and eating disorders. The World Health Organization (WHO) reports that approximately 970 million people worldwide were living with a mental disorder in 2019, with anxiety (301 million) and depressive disorders (280 million) being the most common.[1] The impact of these disorders extends far beyond individual suffering, affecting families, communities, and economies on a global scale. The etiology of mental health disorders is complex and multifaceted, involving an intricate interplay of genetic predisposition, environmental factors, and neurobiological processes. Neurotransmitter dysregulation, particularly involving serotonin, dopamine, and norepinephrine, plays a crucial role in the manifestation of various mental health conditions. While some disorders typically emerge in early adulthood, others can affect individuals across the lifespan, underscoring the need for vigilant monitoring and early intervention strategies.

A. Diagnosis and Current Challenges

The diagnosis of mental health disorders presents unique challenges due to the subjective nature of symptoms and the absence of definitive biological markers. Current diagnostic methods rely heavily on clinical interviews, psychological assessments, and behavioral observations. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[2] and the International Classification of Diseases (ICD-11)[3] provide standardized criteria for diagnosis, but their application requires considerable clinical expertise and time.
Traditional diagnostic approaches often necessitate multiple sessions and extensive data collection to accurately predict mental health conditions. This process can be time-consuming and may delay crucial interventions. Moreover, the overlap of symptoms across different disorders and the influence of cultural and individual factors further complicate the diagnostic process.

B. Research Motivation

The global burden of mental health disorders has reached alarming proportions, with the WHO reporting that mental health conditions account for 13% of the global burden of disease. The COVID-19 pandemic has further exacerbated this crisis, with a 25% increase in the prevalence of anxiety and depression worldwide during the first year of the pandemic.[4] Early detection and intervention are crucial in mitigating the long-term impact of mental health disorders. Studies show that early treatment can lead to better outcomes, reduced chronicity, and lower healthcare costs. However, traditional mental health services face significant challenges in meeting this need, including:

- Shortage of mental health professionals, particularly in low and middle-income countries.
- Stigma associated with seeking mental health care.
- Long waiting times for appointments.

These factors underscore the urgent need for innovative, scalable solutions that can provide early detection of mental health conditions

C. Research Objectives

This study aims to:

1) Comparative Analysis of Deep Learning Models: We conducted an extensive comparative study of various deep learning architectures, including Convolutional Neural Networks (CNN) with Bidirectional LSTM (Bi-LSTM), Text CNN with K-Fold Cross-Validation and BERT (Bidirectional Encoder Representations from Transformers). This analysis provides valuable insights into the efficacy of different models for mental health detection tasks.

2) To evaluate the accuracy and effectiveness of these models in classifying mental health states, including anxiety, depression, suicidal tendencies, stress, and normal.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of related work in the field of deep learning based mental health detection. Section 3 details our methodology, including dataset description, data preprocessing techniques, model architectures, and our proposed framework for real-time mental health detection. Section 4 presents our results and offers an in-depth discussion of our findings, including a comparative analysis of model performance and insights derived from real-time data. Finally, Section 5 concludes the paper, summarizing our key findings and outlining directions for future research in this critical area of mental health informatics.

II. Literature Review

1) Introduction

The application of deep learning techniques in mental health detection has garnered significant attention in recent years. Researchers have explored various architectures and methodologies to improve the accuracy and reliability of AI-driven mental health assessments. This section reviews key studies that have contributed to the advancement of this field.

2) Social Media-Based Detection

Several studies have focused on leveraging social media data for mental health detection. Kholifah et al. (2020) developed a hybrid model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to analyze both text and image data from Twitter. Their model demonstrated high accuracy in identifying depression, anxiety, and bipolar disorder, highlighting the potential of social media mining for early detection of mental health issues.[5].

Kim et al. (2020) utilized a Bidirectional Encoder Representations from Transformers (BERT) architecture to analyze Reddit posts. Their model showed high accuracy in detecting depression, anxiety, and suicidal ideation, emphasizing the effectiveness of transformer-based models in processing social media content for mental health screening.[6]
Husseini Orabi et al. (2018) compared various neural network architectures for detecting depression in Twitter users. Their research found that a hybrid model combining CNNs and Recurrent Neural Networks (RNNs) outperformed other architectures, demonstrating the potential of ensemble approaches in social media-based mental health monitoring.[7]

3) Multimodal Approaches

Zhang et al. (2020) proposed a multimodal deep learning framework that integrated facial expression analysis, speech processing, and text analysis. By fusing CNNs and RNNs, their approach showed improved accuracy in detecting depression and anxiety compared to unimodal methods, suggesting that incorporating multiple data modalities can enhance the robustness of mental health detection systems.[8]

4) Transfer Learning and Pre-trained Models

Ameer et al. (2022) explored the application of transfer learning in mental illness classification on social media texts. They compared various pre-trained language models, including BERT, RoBERTa, and XLNet, and found that fine-tuning these models on mental health-related datasets significantly improved classification accuracy. This study underscores the potential of leveraging large language models for specific mental health detection tasks.performance.[9]

5) Physiological Measures and AI

Shafiei et al. (2020) utilized deep neural networks trained on visual metrics, such as eye movement patterns and pupil responses during cognitive tasks, to identify mental health status. Their model achieved high accuracy in distinguishing between individuals with and without mental health conditions, highlighting the potential of combining objective physiological measures with AI for mental health assessment.[10]

6) Ensemble and Hybrid Approaches

Nasrullah and Jalali (2022) proposed an ensembled deep learning model for detecting types of mental illness through social network analysis. By combining multiple neural network architectures, including CNNs and LSTMs, their approach demonstrated improved accuracy in identifying specific mental health conditions compared to single models, emphasizing the potential of hybrid AI systems in mental health diagnostics.[11]

7) Early Detection and Proactive Intervention

Singh et al. (2023) focused on early-stage depression detection using a novel architecture that combined CNNs with attention mechanisms. Their model showed high sensitivity in identifying subtle indicators of depression onset, underscoring the potential of AI in enabling proactive mental health interventions through early detection.[12]

8) Comparative Studies and Reviews

Several studies have provided comprehensive reviews and comparisons of different AI approaches in mental health detection. Iyortsuun et al. (2023) examined various machine learning and deep learning approaches for mental health diagnosis, finding that deep learning models generally outperformed traditional machine learning methods. Their review highlighted the growing potential of AI in mental health diagnostics while noting challenges in data privacy and model interpretability.[13] Hasib et al. (2023) surveyed machine learning and deep learning techniques for depression detection from social networks. Their analysis revealed that deep learning methods, particularly those using word embeddings and attention mechanisms, showed superior performance. This survey paper provides valuable insights into the evolving landscape of AI-driven depression detection and identifies areas for future research.[14] Bhavani and Naveen (2024) compared various algorithms, including Support Vector Machines, Random Forests, and neural networks, on a diverse dataset of clinical and social media data. Their findings emphasized the importance of feature selection and model optimization in AI-driven mental health assessment, with deep learning models showing superior performance in multi-class classification of mental health disorders.[15] In conclusion, the literature reveals a clear trend towards the use of advanced deep learning techniques, particularly transformer-based models and multimodal approaches, in mental health detection. While these AI-driven methods show great promise in improving the accuracy and early detection of mental health conditions, challenges remain in areas such as data privacy, model interpretability, and clinical integration. Future research should focus on addressing these challenges while continuing to leverage the latest advancements in AI to enhance mental health assessment and intervention strategies.

III. Methodology

A. Dataset Description

For this study, we utilized a comprehensive mental health sentiment dataset obtained from Kaggle. This dataset represents a diverse collection of textual statements sourced from multiple platforms, primarily Reddit and Twitter, as well as several other curated datasets. The amalgamation of these sources provides a robust foundation for training and evaluating deep learning models for mental health detection. Each of the statement is categorized into one of five distinct sentiment classes: normal, anxiety, depression, suicidal, and stress.

B. Data Composition and Statistics

Total Entries: The initial dataset comprised 48964 entries. **Sentiment Classes:** Each entry is categorized into one of five distinct sentiment classes:

- Normal: 16351 entries (33.39% of the dataset)
- Anxiety: 15404 entries (39% of the dataset)
- Depression: 10652 entries (21.75% of the dataset)
- Suicidal: 3888 entries (7.94% of the dataset)
- Stress: 2669 entries (5.45% of the dataset)

C. Data preprocessing and cleaning

data preparation process involved several key steps to ensure the quality and consistency of the textual data for analysis:

Text Preprocessing:

- Converted all text to lowercase.
- Removed punctuation and special characters.
- Eliminated common English stop words.
- Reduced words to their base form through lemmatization.

Data Cleaning:

- Filled missing values in the 'statement' column with empty strings.
- Applied the text preprocessing steps to all entries in the 'statement' column.
- Removed rows with missing values in the 'status' column.
- Filtered out entries with unexpected label values.

Label Encoding:

- Converted sentiment categories to numerical values: Anxiety: 0, Depression: 1, Suicidal: 2, Normal: 3, Stress: 4

D. Prediction Models

In this section, we explore four deep learning models designed for predicting mental health statuses from textual data. The first model combines Convolutional Neural Networks (CNNs) with Bidirectional LSTM layers to capture both local and sequential patterns in the text. The second approach employs a similar CNN architecture but evaluates its performance using K-fold cross-validation to ensure robust and reliable results.The third model utilises the BERT-based model leveraging its pretrained contextual embeddings. Lastly, we utilize a RoBERT-based model, to enhance classification accuracy. Each model's architecture, training methodology, and performance metrics are examined to determine their effectiveness in mental health prediction.

1) Hybrid Model: Integration of CNN and LSTM

Understanding Convolutional Neural Networks (CNN) for Text Classification:

Convolutional Neural Networks (CNNs) are widely recognized for their effectiveness in image processing but have also been successfully applied to text classification tasks. In the context of NLP, CNNs are particularly useful for capturing local patterns in text data, such as phrases or n-grams, which are crucial for understanding the meaning of the text.

CNNs apply a series of filters (convolutional kernels) across the input text sequences. Each filter is responsible for detecting specific features, such as a combination of words that frequently occur together and are indicative of

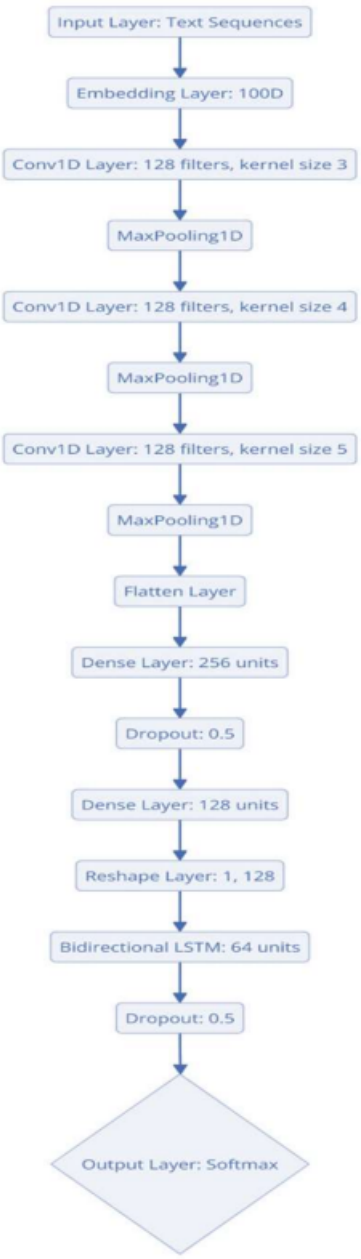

Fig. 1. CNN Bi-directional LSTM Architecture.

a particular class. The convolutional layers output feature maps that highlight the presence of these patterns at various positions in the text.

Following the convolutional layers, max-pooling layers are used to reduce the dimensionality of the feature maps, thereby retaining only the most significant features and reducing computational complexity. The output from these layers is then flattened into a single vector that represents the key features extracted from the entire text input. This vector is passed through fully connected layers for classification.

Long Short-Term Memory (LSTM) Networks for Sequence Modeling:

Long Short-Term Memory (LSTM) networks are a specialized form of Recurrent Neural Networks (RNNs) designed to address the limitations of traditional RNNs, particularly the vanishing gradient problem. LSTMs are well-suited for tasks that involve sequential data, such as text, where the order of words and long-term dependencies between them are crucial.

LSTM networks consist of memory cells that maintain a state over time, allowing the network to retain information from previous time steps and use it to influence future outputs. In the context of text classification, LSTMs can capture the context of a word based on the surrounding words, both preceding and following it. This capability is further enhanced by using a Bi-Directional LSTM (BiLSTM), which processes the input sequence in both directions—forward and backward—capturing dependencies in both directions.

To leverage the strengths of both CNNs and LSTMs, we implemented a hybrid model that combines these two architectures. The CNN component captures local textual patterns, while the LSTM component models the sequential dependencies, providing a comprehensive understanding of the text data.

Model Architecture:

- **Embedding Layer:** The model begins with an embedding layer that converts the input text sequences into dense vector representations. This embedding layer is crucial for translating the discrete word indices into continuous vector spaces where semantically similar words have similar representations.
- **TextCNN Component:** The embedded text is then passed through a series of convolutional layers with varying filter sizes (3, 4, and 5) to capture n-grams of different lengths. Each convolutional layer is followed by a max-pooling layer to down-sample the feature maps, retaining the most salient features while reducing the overall dimensionality.
- **Fully Connected Layers:**After the final convolutional layer, the output is flattened and passed through fully connected layers. These layers further abstract the features extracted by the CNN, enabling the model to learn more complex patterns that may be indicative of specific mental health conditions.
- **BiLSTM Component:** The output from the fully connected layers is reshaped to introduce a time dimension, making it suitable for input into the BiLSTM layer. The BiLSTM processes the sequence in both forward and backward directions, capturing context from the entire text sequence. The output from the BiLSTM layer is then passed through a dropout layer to prevent overfitting.
- **Output Layer:**The final output layer consists of a dense layer with a softmax activation function, which produces

a probability distribution over the four mental health statuses: Anxiety, Depression, Suicidal, and Normal.

Training Strategy

The hybrid model was trained using the Adam optimizer and the sparse categorical cross-entropy loss function, which is appropriate for multi-class classification problems. The model was evaluated on a separate test set to assess its generalization performance. Early stopping was employed during training to prevent overfitting, with the validation loss monitored to restore the best weights.

2) Gated Recurrent Units (GRU) with K-Fold Cross-Validation

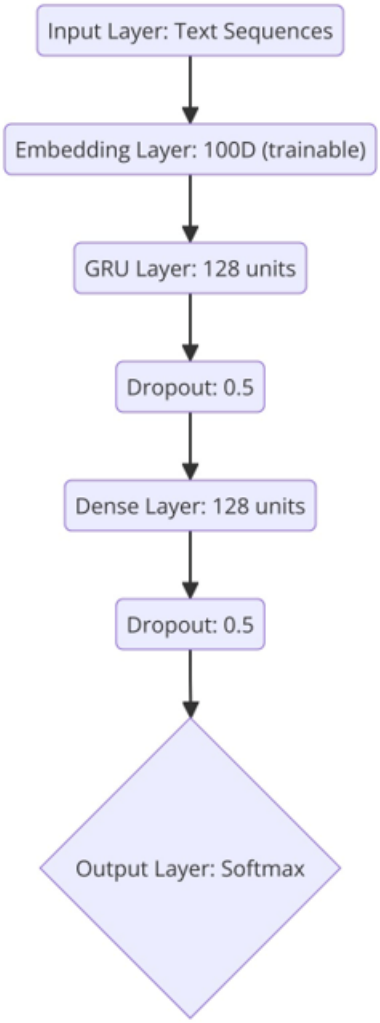

Fig. 2. GRU model Architecture.

Introduction to Gated Recurrent Units (GRU)

Gated Recurrent Units (GRUs) are a variant of Recurrent Neural Networks (RNNs) designed to address some of the limitations of standard RNNs, particularly issues related to long-term dependencies in sequential data. GRUs are similar to Long Short-Term Memory (LSTM) networks but are simpler in design and typically require fewer computational resources, making them an attractive choice for many sequence modeling tasks.

GRUs, like LSTMs, are capable of retaining information over long sequences, which is crucial for tasks such as text classification where context and order of words are important. GRUs achieve this by using gating mechanisms—specifically, the update gate and reset gate—to control the flow of information through the network. These gates help the GRU decide which information to keep or discard as the input sequence is processed, allowing the network to maintain relevant context while reducing the impact of less important information.

Model Architecture

The model used for text classification in this study combines the capabilities of an embedding layer, a GRU layer, and fully connected layers, optimized through K-fold cross-validation. The model architecture is designed as follows:

- **Embedding Layer:** The model starts with an embedding layer that converts the input text into dense vector representations. The embedding layer is initialized randomly and is trainable, allowing the model to learn the most useful representations of words in the context of the given classification task.
- **GRU Layer:** The embedded sequences are then passed through a GRU layer with 128 units. This layer processes the entire sequence and outputs a fixed-size vector that captures the essential sequential dependencies and contextual information from the input text.
- **Fully Connected Layer:** The output of the GRU layer is fed into a fully connected layer with 128 units, followed by a dropout layer with a rate of 0.5 to prevent overfitting. The ReLU activation function is applied in this layer to introduce non-linearity, allowing the model to learn complex patterns in the data.
- **Output Layer:**The final layer of the model is a softmax layer with an output size equal to the number of classes (in this case, five classes corresponding to Anxiety, Depression, Suicidal, Normal, and Stress). This layer outputs the probability distribution over the classes for each input sequence.

Training Strategy: K-Fold Cross-Validation

To evaluate the model's performance more robustly and reduce the risk of overfitting, K-fold cross-validation was implemented. This approach divides the dataset

into 'K' subsets or folds, where the model is trained on 'K-1' folds and validated on the remaining one fold. The process is repeated 'K' times, with each fold serving as the validation set once. The results from each fold are then averaged to obtain a more generalized performance metric.

3) Using BERT(Bidirectional Encoder Representations from Transformers)

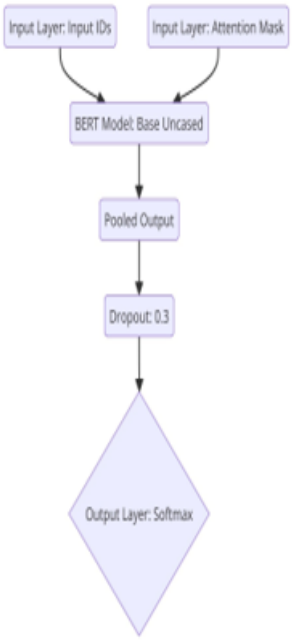

Fig. 3. BERT model Architecture.

Introduction to BERT (Bidirectional Encoder Representations from Transformers)

BERT is a state-of-the-art pretrained model based on the Transformer architecture. Unlike traditional models that process text in a single direction (left-to-right or right-to-left), BERT is bidirectional, meaning it considers context from both directions simultaneously. This bidirectional approach allows BERT to capture deeper and more nuanced meanings of words in their context, making it particularly effective for various natural language understanding tasks, including text classification.

In text classification tasks, BERT is utilized for its ability to generate context-aware embeddings for each token in a sequence. These embeddings are then processed by additional neural network layers to predict the class labels. The pretrained BERT model provides a rich and generalized representation of text, which can be fine-tuned for specific classification tasks, such as identifying mental health statuses from textual data.

Model Architecture:

- **Tokenization and Encoding:** The input text is first tokenized using the BERT tokenizer. This tokenizer converts text into a sequence of token IDs and creates attention masks to indicate which tokens should be attended to. The tokenized inputs are then padded or truncated to ensure a uniform length.
- **BERT Model:**The BERT model processes the tokenized inputs. Specifically, the model generates embeddings for each token, and the pooler_output—which is the output of the [CLS] token—serves as a summary representation of the entire sequence. This pooled output captures the contextualized meaning of the input text.
- **Dropout Layer:** A dropout layer with a rate of 0.3 is applied to the pooled output to prevent overfitting by randomly dropping units during training.
- **Fully Connected Output Layer:** The processed output from the dropout layer is passed through a dense layer with a softmax activation function. This layer produces a probability distribution over the predefined class labels, which in this case are Anxiety, Depression, Suicidal, Normal, and Stress.

Training Strategy

The training process involves fine-tuning the BERT model on the specific classification task.
The model is compiled with the Adam optimizer, a learning rate of 2e-5, and sparse categorical cross-entropy loss. The model is trained for 10 epochs with a batch size of 50, using both the input IDs and attention masks.

4) RoBERTa (Robustly Optimized BERT Approach)

Introduction to RoBERTa (Robustly Optimized BERT Approach)

RoBERTa, introduced by Liu et al., is an advanced language model that builds upon the BERT architecture by incorporating several key improvements. While BERT (Bidirectional Encoder Representations from Transformers) processes text in both directions simultaneously to capture deeper contextual meanings, RoBERTa refines this approach by optimizing the training process and model parameters. These optimizations include training on more data, using dynamic masking, and eliminating the Next Sentence Prediction (NSP) task. As a result, RoBERTa achieves enhanced performance across various natural language understanding tasks, including text classification.

In text classification tasks, RoBERTa excels due to its ability to produce highly contextualized embeddings for each

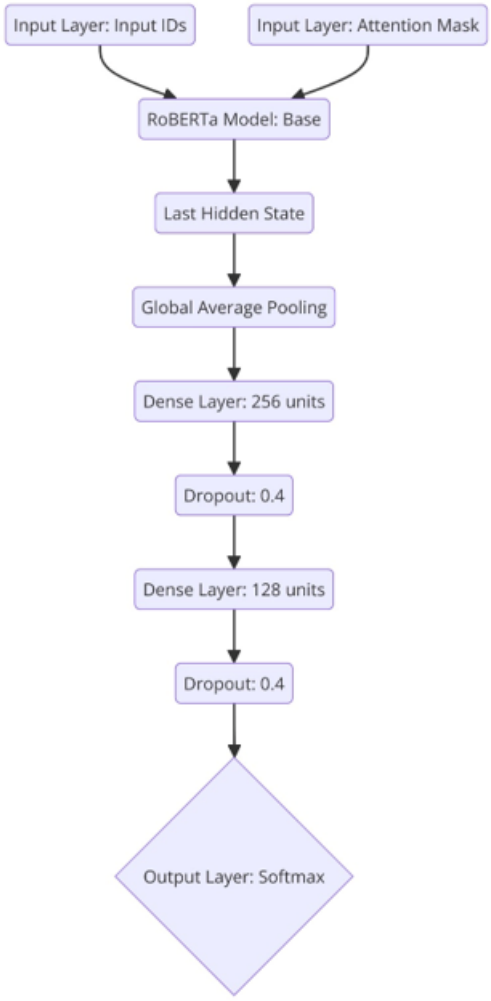

Fig. 4. RoBERT model Architecture.

token in a sequence. These embeddings are then fed into additional neural network layers to predict class labels. The pretrained RoBERTa model provides a robust representation of text that can be fine-tuned for specific classification tasks, such as identifying mental health statuses from text data.

Model Architecture

- **Tokenization and Encoding:** The input text is first tokenized using the RoBERTa tokenizer, which converts text into a sequence of token IDs and generates attention masks to specify which tokens should be considered. The tokenized inputs are then padded or truncated to a uniform length of 128 tokens.
- **RoBERTa Model:** The RoBERTa model processes the tokenized inputs to generate contextual embeddings. Specifically, the model produces a sequence of hidden states, from which the final hidden state of the sequence is pooled using global average pooling. This pooled output provides a summary representation of the entire input text.
- **Dropout Layers:** Two dropout layers with rates of 0.4 are applied to the pooled output to mitigate overfitting by randomly dropping units during training.
- **Fully Connected Output Layer:** The processed output is passed through a dense layer with a softmax activation function. This layer generates a probability distribution over the predefined class labels, including Anxiety, Depression, Suicidal, Normal, and Stress.

Training Strategy

The model is compiled with the Adam optimizer, a learning rate of 3e-5, and sparse categorical cross-entropy loss. Training is conducted for 10 epochs with a batch size of 50, using both input IDs and attention masks.

IV. Evaluations and Results

TABLE I
Precision, Recall and F1 score for CNN Bi-directional LSTM

Label	Precision	Recall	F1-Score	Support
Anxiety	0.35	0.56	0.43	784
Depression	0.56	0.67	0.61	3001
Suicidal	0.58	0.47	0.52	2079
Normal	0.89	0.87	0.88	3379
Stress	0.00	0.00	0.00	550
Accuracy			0.65	9793
Macro Avg	0.48	0.51	0.49	9793
Weighted Avg	0.63	0.65	0.64	9793

The results for the Text CNN with Bi-directional LSTM model, as presented in Table I, demonstrate varied performance across different mental health categories. The model achieved an overall accuracy of 65%, with a weighted average F1-score of 0.64. Notably, the model performed exceptionally well in identifying ”Normal” instances, with a high precision of 0.89 and an F1-score of 0.88. This category also had the highest support with 3379 samples, indicating a well-represented class in the dataset. However, the model struggled significantly with the ”Stress” category, where it failed to correctly identify any instances, resulting in precision, recall, and F1-score all at 0.00. This poor performance on the ”Stress” category, despite having 550 samples, suggests a critical area for improvement in the model’s ability to detect stress-related mental health issues. The ”Depression” and ”Suicidal” categories showed moderate performance, with F1-scores of 0.61 and 0.52, respectively. For depression, the model achieved a better recall (0.67) than precision (0.56), indicating a tendency to over-predict this category. Conversely, for suicidal tendencies, the model had higher precision (0.58) than recall (0.47), suggesting it might miss some cases but is more accurate when it does predict this category. The model’s ability to detect ”Anxiety” was limited, achieving an F1-score of 0.43, with a notably low precision of 0.35 but a higher recall of

0.56. This imbalance suggests that while the model identifies many anxiety cases, it also incorrectly classifies other mental health states as anxiety. Overall, while the Text CNN with Bi-directional LSTM model was effective in certain categories, particularly "Normal," there is significant room for improvement in accurately classifying more complex and subtle mental health conditions like "Stress" and "Anxiety." The model's performance highlights the challenges in multi-class classification of mental health states and underscores the need for further refinement of the model architecture or training approach, especially for underperforming categories.

TABLE II
PRECISION, RECALL AND F1 SCORE FOR GRU MODEL

Label	Precision	Recall	F1-Score	Support
Anxiety	0.73	0.76	0.75	766
Depression	0.71	0.69	0.70	3108
Suicidal	0.62	0.66	0.64	2131
Normal	0.89	0.91	0.90	3238
Stress	0.64	0.45	0.53	549
Accuracy			0.75	9792
Macro Avg	0.72	0.69	0.70	9792
Weighted Avg	0.75	0.75	0.75	9792

The results for the GRU model with k-fold cross-validation, as presented in Table II, demonstrate a notable improvement in performance across all mental health categories compared to the previous CNN Bi-directional LSTM model. The GRU model achieved an overall accuracy of 75%, with a consistent weighted average F1-score of 0.75, indicating balanced performance across classes. The model performed exceptionally well in identifying "Normal" instances, with a high precision of 0.89 and an impressive F1-score of 0.90. This category also had the highest support with 3238 samples, suggesting robust performance on the most prevalent class in the dataset. Notably, the GRU model showed significant improvement in detecting "Anxiety," achieving an F1-score of 0.75, with both high precision (0.73) and recall (0.76). This marks a substantial enhancement over the previous model's performance in this category. The model's performance on "Depression" and "Suicidal" categories was also strong, with F1-scores of 0.70 and 0.64 respectively. For depression, the model achieved a balanced precision (0.71) and recall (0.69), indicating consistent performance. The suicidal category showed slightly lower but still substantial scores, with a precision of 0.62 and recall of 0.66. Importantly, unlike the previous model, the GRU model demonstrated an ability to detect "Stress" cases, achieving an F1-score of 0.53. While this is the lowest among all categories, it represents a significant improvement from the previous model's complete failure in this category. Overall, the GRU model with k-fold cross-validation shows a more balanced and improved performance across all mental health categories. The consistent accuracy and F1-scores across macro and weighted averages (0.70 and 0.75 respectively) suggest that the model performs well even with class imbalances. This model appears to be more robust in handling the complexities of multi-class mental health classification, though there is still room for improvement, particularly in the "Stress" category.

TABLE III
PRECISION, RECALL AND F1 SCORE FOR BERT MODEL

Label	Precision	Recall	F1-Score	Support
Anxiety	0.81	0.92	0.86	784
Depression	0.74	0.78	0.72	3001
Suicidal	0.70	0.61	0.66	2079
Normal	0.93	0.94	0.93	3379
Stress	0.69	0.65	0.67	550
Accuracy			0.80	9793
Macro Avg	0.77	0.78	0.77	9793
Weighted Avg	0.80	0.80	0.80	9793

The results for the BERT model, as presented in Table III, demonstrate exceptional performance across all mental health categories, further improving upon the already strong results of the previous models. The BERT model achieved an outstanding overall accuracy of 80%, with a consistent weighted average F1-score of 0.80, indicating robust and balanced performance across classes. The model excelled particularly in identifying "Normal" instances, achieving an exceptionally high precision of 0.93 and recall of 0.94, resulting in an F1-score of 0.93. This category also had the highest support with 3379 samples, suggesting excellent performance on the most prevalent class in the dataset. Notably, the BERT model showed remarkable performance in detecting "Anxiety," achieving the highest F1-score among all categories at 0.86, with both high precision (0.81) and recall (0.92). This represents a substantial improvement over previous models and indicates BERT's strong capability in identifying anxiety-related text patterns with high sensitivity. The model's performance on "Depression" and "Suicidal" categories was also impressive, with F1-scores of 0.76 and 0.66 respectively. For depression, the model showed balanced precision (0.74) and recall (0.78), indicating consistent performance. The suicidal category had lower but still substantial scores, with a precision of 0.70 and recall of 0.61, suggesting room for improvement in this critical category. Interestingly, the BERT model's performance on the "Stress" category, while improved from some previous models, still shows room for enhancement with an F1-score of 0.67. The model demonstrated higher recall (0.65) than precision (0.69) for stress detection, indicating a tendency to over-identify stress cases. Overall, the BERT model showcases superior and well-balanced performance across all mental health categories. The high and consistent accuracy, along with strong F1-scores across macro and weighted averages (both at 0.80), suggest that the model performs exceptionally well even with class imbalances. This model appears to be highly effective in handling the complexities of multi-class mental health classification, demonstrating BERT's power in understanding and categorizing nuanced textual information related to mental health states. The results highlight BERT's potential as a powerful tool for automated mental health assessment, though

continued refinement, especially for categories like "Suicidal" and "Stress," could further enhance its clinical applicability.

TABLE IV
PRECISION, RECALL AND F1 SCORE FOR ROBERT MODEL

Label	Precision	Recall	F1-Score	Support
Anxiety	0.82	0.85	0.84	784
Depression	0.84	0.64	0.72	3001
Suicidal	0.64	0.83	0.72	2079
Normal	0.94	0.93	0.93	3379
Stress	0.62	0.74	0.67	550
Accuracy			0.80	9793
Macro Avg	0.77	0.80	0.78	9793
Weighted Avg	0.82	0.80	0.80	9793

The results for the RoBERT model, as presented in Table IV, demonstrate a significant improvement in performance across all mental health categories compared to the previous models. The RoBERT model achieved an impressive overall accuracy of 80%, with a weighted average F1-score of 0.80, indicating robust and balanced performance across classes.

The model excelled in identifying "Normal" instances, achieving a remarkably high precision of 0.94 and an F1-score of 0.93. This category also had the highest support with 3379 samples, suggesting excellent performance on the most prevalent class in the dataset.

Notably, the RoBERT model showed outstanding performance in detecting "Anxiety," achieving the highest F1-score among all categories at 0.84, with both high precision (0.82) and recall (0.85). This represents a substantial improvement over previous models and indicates RoBERT's strong capability in identifying anxiety-related text patterns.

The model's performance on "Depression" and "Suicidal" categories was also impressive, both achieving F1-scores of 0.72. For depression, the model showed high precision (0.84) but lower recall (0.64), suggesting it's more cautious in labeling text as depressive. Conversely, for suicidal tendencies, the model had lower precision (0.64) but high recall (0.83), indicating it's more sensitive in detecting potential suicidal content.

The RoBERT model also demonstrated improved ability to detect "Stress" cases, achieving an F1-score of 0.67. While this is the lowest among all categories, it represents a significant improvement from previous models, with a balanced precision (0.62) and recall (0.74).

Overall, the RoBERT model shows superior and more balanced performance across all mental health categories compared to the previous models. The high and consistent accuracy, along with strong F1-scores across macro and weighted averages (0.78 and 0.80 respectively), suggest that the model performs exceptionally well even with class imbalances. This model appears to be highly effective in handling the complexities of multi-class mental health classification, demonstrating RoBERT's power in understanding and categorizing nuanced textual information related to mental health states.

TABLE V
SUMMARY TABLE: MODEL COMPARISON

Model	Precision	Recall	F1-Score	Accuracy
CNN Bi-directional LSTM	0.63	0.65	0.64	0.65
GRU	0.75	0.75	0.75	0.75
BERT	0.80	0.80	0.80	**0.80**
RoBERT	0.82	0.80	0.80	**0.80**

Our comparative analysis of four different models—CNN Bi-directional LSTM, GRU, RoBERT, and BERT—revealed varying levels of performance in classifying mental health conditions. The results, summarized in Table V, indicate that the RoBERT and BERT models demonstrated superior performance, both achieving identical overall metrics with a weighted average precision of 0.80-0.82, recall of 0.80, F1-score of 0.80, and accuracy of 0.80. The GRU model showed competitive performance, with all metrics at 0.75, positioning it as a strong alternative. In contrast, the CNN Bi-directional LSTM model exhibited the lowest performance among the four, with a weighted average precision of 0.63, recall of 0.65, F1-score of 0.64, and accuracy of 0.65. Notably, all models maintained a balance between precision and recall, as evidenced by F1-scores closely aligning with the average of precision and recall. These findings suggest that transformer-based models (RoBERT and BERT) outperform traditional recurrent neural networks in this mental health classification task, potentially due to their advanced contextual understanding capabilities.

V. CONCLUSION

This study presents a comprehensive analysis of four advanced deep learning approaches for text-based mental health classification: a TextCNN model integrated with bi-directional LSTM, a GRU model utilizing K-fold cross-validation, a fine-tuned BERT-based model, and a RoBERTa model. Our comparative analysis reveals significant insights into the efficacy of these models in classifying mental health states including anxiety, depression, suicidal tendencies, stress, and normal conditions.

Key findings include:

1) Transformer-based models (BERT and RoBERTa) demonstrated superior performance, both achieving an overall accuracy of 80% and a weighted average F1-score of 0.80. This suggests that these models' advanced contextual understanding capabilities are particularly effective for mental health classification tasks.
2) The GRU model with K-fold cross-validation showed competitive performance, with an accuracy of 75% and consistent F1-scores across categories, indicating its robustness in handling class imbalances.
3) The CNN Bi-directional LSTM model, while less effective than the other models, still achieved a respectable

65% accuracy, highlighting the potential of hybrid architectures in this domain.

4) All models showed varying performance across different mental health categories, with generally high accuracy in identifying "Normal" states but challenges in distinguishing between more nuanced conditions like "Stress" and "Anxiety".

Future Works:

1) Multi-modal analysis: Integrate additional data sources such as speech patterns, facial expressions, or physiological markers to create a more comprehensive mental health assessment system.
2) Longitudinal studies: Implement these models in long-term studies to evaluate their effectiveness in tracking mental health changes over time and predicting potential mental health crises.
3) Test and adapt these models across different languages and cultural contexts to ensure their applicability in diverse populations.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders
[2] https://www.psychiatry.org/psychiatrists/practice/dsm
[3] https://icdcdn.who.int/icd11referenceguide/en/html/index.html
[4] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide
[5] Kholifah, B., Syarif, I., & Badriyah, T. (2020). Mental Disorder Detection via Social Media Mining using Deep Learning. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 5(4).
[6] Kim, J., Lee, J., Park, E. et al. A deep learning model for detecting mental illness from user content on social media. Sci Rep 10, 11846 (2020).
[7] Deep Learning for Depression Detection of Twitter Users (https://aclanthology.org/W18-0609) (Husseini Orabi et al., CLPsych 2018
[8] Z. Zhang, W. Lin, M. Liu and M. Mahmoud, "Multimodal Deep Learning Framework for Mental Disorder Recognition," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp. 344-350, doi: 10.1109/FG47880.2020.00033.
[9] Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, Alexander Gelbukh. Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning. doi: https://arxiv.org/abs/2207.01012
[10] Shafiei, S.B., Lone, Z., Elsayed, A.S. et al. Identifying mental health status using deep neural network trained by visual metrics. Transl Psychiatry 10, 430 (2020). https://doi.org/10.1038/s41398-020-01117-5
[11] Syed Nasrullah, Asadullah Jalali. Detection of Types of Mental Illness through the Social Network Using Ensembled Deep Learning Model. doi: https://doi.org/10.1155/2022/9404242
[12] Singh, K., Ahirwal, M.K. & Pandey, M. Mental Health Monitoring Using Deep Learning Technique for Early-Stage Depression Detection. SN COMPUT. SCI. 4, 701 (2023).
[13] Iyortsuun, N.K.; Kim, S.-H.; Jhon, M.; Yang, H.-J.; Pant, S. A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. Healthcare 2023, 11, 285. https://doi.org/10.3390/healthcare11030285
[14] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak and M. S. Alam, "Depression Detection From Social Networks Data Based on Machine Learning and Deep Learning Techniques: An Interrogative Survey," in IEEE Transactions on Computational Social Systems, vol. 10, no. 4, pp. 1568-1586, Aug. 2023, doi: 10.1109/TCSS.2023.3263128
[15] B. H. Bhavani and N. C. Naveen, "An Approach to Determine and Categorize Mental Health Condition using Machine Learning and Deep Learning Models", Eng. Technol. Appl. Sci. Res., vol. 14, no. 2, pp. 13780–13786, Apr. 2024.