Seventh International Conference in Information and Communication Technologies,
Lusaka, Zambia 15th to 16th October 2025

Using XAI and Visualization for Clinical Decision-Making: Targeting Mental Health Assessment Using Social Media Data

1st Chekani Chiume
School Of Information and Communication Technology
Copperbelt University
Kitwe,Zambia chekanichiume@gmail.com

2nd Maybins Lengwe
School Of Information and Communication Technology
Copperbelt University
Kitwe,Zambia
kasalwelengwe@gmail.com

3rd Calvin Silwizya
School Of Information and Communication Technology
Zambia University College Of Technology
Ndola, Zambia
silwizyacs@zut.edu.zm

Abstract—Depression is a major global mental health challenge, often going underdiagnosed due to stigma, delayed recognition, and limited monitoring, reducing opportunities for timely intervention. This study applied Explainable Artificial Intelligence (XAI) and visualization techniques to detect signs of depression from social media data, aiming to improve both predictive accuracy and interpretability. A dataset of 27,977 anonymized public posts was collected from X (formerly Twitter) between 2023 and 2024, cleaned and preprocessed to yield 26,925 posts. Two models were developed: a TF-IDF + Random Forest baseline and a fine-tuned DistilBERT transformer model, trained and evaluated using an 80/10/10 train–validation–test split. DistilBERT outperformed the baseline, achieving 92% accuracy, 89% precision, 91% recall, and a 0.94 ROC-AUC score. Error analysis revealed that false positives often involved sarcasm, while false negatives reflected subtle or metaphorical distress. To enhance interpretability, SHAP (Shapley Additive Explanations) was used for global feature importance, while LIME (Local Interpretable Model-Agnostic Explanations) provided local, case-level insights. Validation by mental health professionals confirmed that 86% of explanations aligned with clinical indicators. Visualization tools, including SHAP plots and saliency maps, further improved accessibility of results. Ethical safeguards such as data anonymization and compliance with platform policies were enforced. This work demonstrates that combining transformer-based models with XAI can produce accurate, interpretable frameworks that bridge the gap between AI prediction and clinical reasoning, supporting responsible and trustworthy mental health assessment.

Index Terms—Depression detection, Explainable Artificial Intelligence (XAI), Natural Language Processing (NLP), Transformer models, SHAP, LIME, Visualization, Mental health assessment.

I. Introduction

Depression is a leading cause of disability worldwide, affecting over 280 million people, according to the World Health Organization (WHO) [1]. Despite its prevalence, depression is often underdiagnosed due to social stigma, limited access to mental health services, delayed diagnosis, and insufficient continuous monitoring [2]. Traditional assessments, such as clinical interviews and self-reported questionnaires, provide structured insights but may miss early warning signs, especially in individuals reluctant to seek help or those with fluctuating symptoms, thereby limiting timely intervention [3]. Recent advances in artificial intelligence (AI) offer new opportunities to complement traditional methods by analyzing digital footprints from social media, where users frequently express emotions and share personal experiences. Machine learning (ML) models trained on such data demonstrate strong predictive performance. However, these models often function as "black boxes," meaning their internal decision-making processes are complex and not easily understood, which limits their acceptance in healthcare, where transparency, interpretability, and trust are essential for ethical use [4]. Moreover, without effective visualization and clear explanations of how decisions are made, these complex models remain difficult to interpret, further hindering their adoption in healthcare settings. Even interpretable models may fail to deliver actionable insights to clinicians, thereby restricting their clinical utility. This study proposes a machine learning framework that combines explainable artificial intelligence (XAI) techniques with visualization to detect depression from social media data. Our approach consists of two phases: (1) baseline experiments using traditional natural language processing (NLP) methods including tokenization, stemming, and term frequency–inverse document frequency (TF-IDF) with Random Forest classifiers; and (2) applying a fine-tuned Distil BERT model a compact, efficient variant of Bidirectional Encoder Representations from Transformers (BERT) to capture deeper semantic and contextual features [5]. To enhance interpretability and clinician trust, the study employed two widely used XAI methods: SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These techniques revealed key predictive features, including the use of negative emotion words, frequent use of first-person pronouns, and expressions of loneliness and hopelessness. Visualization techniques were integrated to present model explanations in a clear, user-friendly format that supported clinical decisionmaking. Our experiments demonstrate that the Distil BERT model outperforms the baseline, achieving 92% accuracy, 89% precision, 91% recall, and a 90% F1- Score. The rest of this paper is organized as follows: Section II reviews related work on depression

 ISBN: 978-9982-95-500-3

detection, explainable artificial intelligence, and visualization in healthcare; Section III details the methodology, including data collection, pre-processing, model development, evaluation metrics, and explanation techniques; Section IV presents experimental results; Section V discusses findings, implications, and limitations; and Section VI concludes and suggests future work.

II. LITERATURE REVIEW

This section presents a review of the existing literature on the application of Explainable Artificial Intelligence (XAI) in mental health systems, with a particular focus on assessment and evaluation. According to [6], XAI techniques have been employed to enhance the transparency of deep learning models, addressing the interpretability challenges commonly associated with black-box systems. This paper also emphasizes the importance of transparency, particularly in mental health contexts, where improved interpretability supports informed clinical decision-making and fosters trust in AI tools. Furthermore, the study conducted by [7] indicated that integrating visualization methods with XAI approaches provides visual representations of the factors influencing model predictions. These techniques aid in the interpretation of complex model outputs and render predictions clinically meaningful. Overall, the reviewed literature suggests that the use of XAI can improve the usability and acceptance of AI-driven mental health assessment systems by making their decision-making processes more transparent and explainable in clinical settings. In this study, papers were reviewed concerning the application of Explainable Artificial Intelligence in mental health, visualization techniques for mental health conditions, and developing Explainable Artificial Intelligence (XAI) models in mental health systems, focusing on enhancing transparency, accuracy, precision, and interpretability to improve the understandability of AI-driven assessments and decisions.

A. Application of Explainable AI in Mental Health

The study [8] employed Explainable Artificial Intelligence (XAI) techniques to enhance the interpretability of a deep learning framework for detecting suicidal ideation. The author demonstrated two phases: classifying social media posts into suicidal and non-suicidal categories and extracting factors contributing to suicidal ideation. The interpretable insights facilitated the identification of risk factors associated with suicidal ideation, thereby improving model transparency and enabling practical application in mental health monitoring. This study showed that integrating XAI with deep learning models can yield accurate and explainable results, which are critical for healthcare domains. In [9], the author explored the application of Explainable Artificial Intelligence (XAI) and deep learning techniques for mental health classification using text-based data. The research focused on identifying key mental health conditions, including addiction, alcoholism, anxiety, depression, and suicidal thoughts, through the analysis of usergenerated content. The methodology employed various text vectorization techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF), Word to Vector (Word2Vec), and Global Vectors for Word Representation (GloVe), to transform unstructured text into machine-readable formats. Emphasizing model transparency, the study utilized Explainable Artificial Intelligence (XAI) techniques, providing interpretable insights into the contributions of features to model predictions. The findings showed that combining traditional machine learning models with explainability tools effectively identified patterns and indicators of mental health issues. The authors [10] examined the application of Explainable Artificial Intelligence (XAI) and deep learning methods for detecting mental health disorders using text-based data collected from social media platforms. The researchers implemented a dual-model framework that compared the performance and interpretability of Bidirectional Long Short-Term Memory (BiLSTM) models with those of transformer-based models pre-trained for mental health-related tasks. To ensure transparency in the model decision-making process, the study used two explanation techniques: Local Interpretable Model-Agnostic Explanations (LIME) and Attention Gradient (AGRAD), a model-specific self-explanation method. The results highlighted a clear tradeoff between prediction accuracy and interpretability, particularly in the context of mental health classification. The study emphasized the importance of transparent machine learning models in clinical and high-stakes environments, where understanding model behavior is crucial for establishing trust and facilitating practical implementation. The reviewed studies highlighted notable advancements in combining Explainable Artificial Intelligence (XAI) with deep learning methods to identify mental health conditions through the analysis of social media text. They noted that techniques like Local Interpretable Model-Agnostic Explanations (LIME) and Attention Gradient (AGRAD) contributed to improving model interpretability while maintaining acceptable accuracy levels. The research further highlighted that understanding how models arrive at decisions is crucial, particularly in clinical environments where ethical and practical considerations are paramount. However, the existing literature lacked the application of predictive analysis. Therefore, this paper addresses the critical need to balance predictive accuracy, model transparency, and precision.

B. Application of Visualization Techniques for Mental Health Conditions

The study [11] introduced a framework for detecting depression through a hybrid visualization approach that integrated both local and global interpretability techniques, enhancing the understanding and adaptation of models. This fusion of statistical modeling and visualization techniques

 ISBN: 978-9982-95-500-3

facilitated an extensive evaluation of model performance and appropriateness across various healthcare settings. The research by [12] employed visualization techniques within Explainable Artificial Intelligence (XAI) frameworks to enhance model interpretability and support clinical decision-making. The authors illustrated how tools such as heatmaps and saliency maps are employed in pathology to emphasize critical features in digital tissue slide images. These visual aids enabled pathologists to identify inconsistencies and refine their diagnostic processes, thereby enhancing diagnostic accuracy and patient outcomes. The study highlighted that visualizationdriven interpretability fosters clinicians' trust in AI-assisted decisions. In another investigation, [10] aimed at merging data visualization with XAI to convert complex AI-generated data into understandable patterns for decision-makers. The authors observed that traditional visualization systems often exhibit inefficiencies, such as high costs and slow response times. To tackle these challenges, they implemented XAIbased visualization tools to enhance the interpretability and practicality of complex models. This integration proved that visualization not only improves model transparency but also facilitates timely and informed decision-making. Furthermore, [13] explored the use of visual analytics (VA) to interpret black-box machine learning models, with a particular focus on deep neural networks. The study demonstrated that integrating VA with XAI enabled users, even those with minimal technical backgrounds, to explore the AI decision-making process interactively. This interactivity bolstered users' confidence in the system and supported enhanced decision-making. The visual analytics methodology thus demonstrated effectiveness in elucidating the inner workings of complicated AI models. The paper [12] underscored the essential role that XAI and visualization techniques play in addressing the transparency challenges associated with deep learning models in healthcare. The research demonstrated the use of interpretability methods, such as LIME and SHAP, alongside visual tools like saliency maps and heat maps, providing vital insights into the rationale behind AI predictions, particularly in medical diagnostics. The results validated that interpretability is essential for the deployment of reliable AI in sensitive healthcare environments, where transparency in decision-making is critical. This study focused on applying XAI and visualization methods to mental health assessments using social media data, aiming to enhance the interpretability and applicability of AI predictions in decision-making. Utilizing techniques such as LIME and SHAP, the study identified key features, including emotional tone, sentiment, and specific keywords, that influenced the model's outputs. Visualization tools were utilized to present these results effectively, enabling users to understand why certain posts were identified as indicators of mental health issues. This strategy improved transparency, making it easier for non-technical users to trust and act on the AI findings. Overall, the study reinforced that the combination of XAI and visualization fosters a better understanding, enhances trust, and makes AI more applicable in real-world mental health settings for monitoring.

III. RESEARCH METHODOLOGY

This study employed an experimental research design to evaluate the effectiveness of machine learning models in detecting mental health indicators using social media data. The dataset comprised 27,977 user-generated posts sourced from the social media platform X, collected from publicly available repositories on Kaggle. Kaggle is a well-known provider of datasets often used in academic research, offering a high level of accessibility and recognized credibility [14]. However, as the data was derived from user-generated content, it required meticulous preparation to ensure consistency and suitability for analysis. The dataset featured posts created from 2023 to 2024. Before model development, the data underwent an extensive cleaning process that addressed missing values, eliminated duplicate entries, corrected formatting issues, and standardized text formats. Pandas was utilized for managing missing and null values, while Numpy facilitated efficient numerical operations. The text data was further pre-processed through normalization steps, which included converting the text to lowercase, removing special characters, eliminating stop words, and performing tokenization. This thorough cleaning procedure helped minimize noise and potential biases, ensuring that the dataset met the necessary quality standards for subsequent analysis and model training.

A. Data Collection and Pre-processing

A dataset featuring 27,977 user-generated posts from the social media platform X was collected from publicly accessible open-source repositories, specifically from a collection of posts on Kaggle [13]. This dataset concentrates on posts that express personal emotions, mental states, or overall well-being, aligning to identify mental health risks. All data underwent anonymization, ensuring that no personally identifiable information (PII) was retained or processed, fully complying with X's platform policies and established ethical research standards. The labelling process utilized an automated, keywordbased method. Lists of keywords associated with depression, anxiety, and other mental health conditions were created based on previous research and expert input. These keyword lists were then applied to assign binary labels to each post: highrisk (indicating the presence of mental health issues) or lowrisk (indicating the absence of such concerns). To ensure label quality, a random sample was manually reviewed to verify the validity of the automated annotations, thereby strengthening the dataset's reliability.

 ISBN: 978-9982-95-500-3

B. Data pre-processing

Data pre-processing is a vital phase within the machine learning pipeline, where raw, unstructured data is transformed into a clean, structured format suitable for analysis and model development. This process enhances data quality, ensures consistency, and boosts usability by tackling noise, correcting inconsistencies, and standardizing formats. During the data cleaning phase, Python tools were employed. Pandas offers functionalities for managing missing values, detecting and removing duplicate entries, and implementing necessary transformations. NumPy was utilized to effectively manage numerical data and conduct array-based operations, optimizing computational efficiency. In addition to cleaning, data set validation was carried out to ensure the integrity and reliability of the data before training. This involved confirming data types, identifying out-of-range or invalid values, and checking for logical consistency across various features, for instance, verifying chronological order in timestamp fields and ensuring that categorical variables corresponded to the expected set of labels. Descriptive statistics and visualizations were used to identify potential anomalies, imbalances, or unusual patterns that could impact model performance. The dataset was meticulously divided into training, validation, and test sets to prevent data leakage and ensure that the model was evaluated on representative, unseen data. These validation procedures were crucial in enhancing the model's accuracy, generalization, and robustness while minimizing the risk of biased or unreliable results. For this study, the data pre-processing regimen consisted of a series of structured operations:

I. Noise Removal: Irrelevant or unnecessary characters,such as special symbols, Uniform Resource Locators (URLs), hashtags, and user mentions, were systematically eliminated due to their minimal contribution to the classification goal.

II. Handling Missing or Incomplete Entries: Incomplete orcorrupt data points were identified and either removed or filled in (imputed) to maintain the integrity of the training dataset.

III. Text Standardization: All textual data was converted tolowercase for consistency. Redundant punctuation was discarded, and tokenization was applied to break down the text into distinct, meaningful units.

C. Machine Learning Models and Explainability Techniques

This section describes the machine learning models and Explainable Artificial Intelligence (XAI) strategies employed for classifying text data related to mental health. We began by establishing a baseline using traditional Natural Language Processing (NLP) techniques, then transitioned to an advanced deep learning model. Finally, we integrated XAI methods to enhance interpretability and transparency. Our approach to classifying text data linked to mental health encompassed two distinct phases: first, setting up a baseline model for foundational insights, and second, deploying a sophisticated deep learning model to tackle more complex contextual issues details.

D. Baseline Traditional Model: Random Forest with TF-IDF

At first, we established a traditional NLP pipeline as our starting point. We employed the Term Frequency-Inverse Document Frequency (TF-IDF) method for feature extraction. TF-IDF converts raw text into weighted numerical forms, efficiently highlighting the significance of words throughout the dataset. This method prepares text for machine learning models by measuring the relevance of each word to a document within a larger collection. Next, we trained a Random Forest classifier using the TF-IDF features. Random Forest is an ensemble learning approach that generates multiple decision trees during the training phase and classifies based on the majority vote of these trees. It's acknowledged for its strength and reliable performance across varied datasets. The model's goal was binary classification, predicting whether a text indicated a high-risk mental health condition. Regarding performance, the Random Forest classifier, functioning as our baseline, recorded an accuracy of 85%, a precision of 82%, a recall of 80%, and an F1-score of 81%. While it offered a strong initial performance, it fell short in capturing the subtle contextual nuances often found in discussions of mental health.

E. Advanced Deep Learning Model: DistilBERT

To enhance classification performance and improve contextual understanding, we utilized DistilBERT (Distilled Bidirectional Encoder Representations from Transformers). DistilBERT is a transformer-based deep learning architecture that offers a lighter and faster alternative to the original BERT model. It retains most of BERT's strong language comprehension abilities while being more efficient in computation. Similar to BERT, DistilBERT is pre-trained on extensive text data to capture general language patterns, enabling it to comprehend intricate semantic relationships. The pre-trained DistilBERT model was then fine-tuned on our targeted labelled mental health dataset. This fine-tuning process involved modifying the model's classification layers for binary predictions, enabling it to specialize in recognizing and understanding the subtle semantic relationships and contextual cues inherent in mental health-related text. This adjustment was essential for progressing beyond mere keyword matching to grasping the deeper meanings within the text. Model Results: The finetuned DistilBERT model significantly surpassed the baseline, attaining an accuracy of 92%, precision of 89%, recall of 91%, and an F1-score of 90%. Its capacity to capture complex semantic relationships resulted in a significant improvement in identifying high-risk mental health conditions compared to traditional methods.

 ISBN: 978-9982-95-500-3

F. Explainable Artificial Intelligence (XAI) Methods

To promote transparency and understanding of our machine learning models, particularly the complex DistilBERT model, we have incorporated Explainable Artificial Intelligence (XAI) techniques. These approaches were essential for elucidating the decision-making processes of our models, enhancing clinician trust and facilitating more precise mental health diagnostics.

G. SHAP (Shapley Additive explanations)

SHAP is a game-theoretic method that quantifies the contribution of each input feature to a model's predictions. By using Shapley values, rooted in cooperative game theory, SHAP fairly allocates the "pay-out" (the prediction's deviation from a baseline) to the "players" (input features). This approach enables both global interpretability, which highlights the significance of features across the entire dataset, and local interpretability, which clarifies individual predictions. We implemented SHAP to pinpoint the terms or expressions that significantly impacted the model's high-risk classifications in particular text samples. By consolidating these findings, SHAP also highlighted the key linguistic and semantic features influencing our model's predictions across the entire dataset. This delivered vital insights into the foundational logic of our deep learning model.

H. LIME (Local Interpretable Model-Agnostic Explanations)

LIME was used to produce local, comprehensible explanations for individual predictions. It functions by generating perturbed versions of the input sample and analysing how the outputs of complex "black-box" models vary. Through these analyses, LIME develops a simpler, interpretable surrogate model (such as a linear model) that mimics the behaviour of the complex classifier around that specific prediction. This process enhances the transparency of individual predictions, enabling us to visualize and understand which words or phrases had a significant influence on a particular outcome. Consequently, it improves clarity and bolsters clinicians' understanding and trust in AI-driven mental health diagnostics. The proposed framework begins with gathering posts from X via web scraping. The acquired data is then cleaned to eliminate noise, including HTML tags, non-numeric characters, and any missing values. Once cleaned, features are extracted using

TF-IDF and DistilBERT techniques. These features are utilized to train both machine learning and deep learning models. Finally, Explainable AI methods, such as SHAP and LIME, are employed to interpret the model outputs, aiding in the identification of mental health risks

I. Model Evaluation

To thoroughly assess model performance, a range of standard classification metrics, including Precision, Recall, and F1Score, were employed. These metrics are especially effective for binary classification tasks, where the repercussions of false positives and false negatives are particularly significant, such as in mental health diagnostics.

J. Precision

Precision measures the proportion of correctly predicted positive cases (True Positives) out of all cases that the model predicted as positive (True Positives + False Positives). It quantifies the model's ability to avoid false alarms, which is crucial in minimizing unnecessary interventions in clinical settings.

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

Where:

TP: True Positives — correctly predicted high-risk cases.

FP: False Positives — low-risk cases incorrectly predicted as high-risk.

High precision indicates that when the model predicts a patient is at high risk, it is very likely to be correct, minimizing false alarms.

K. Recall (Sensitivity)

Recall, often referred to as sensitivity, gauges the percentage of real positive cases accurately identified by the model. This metric is crucial in healthcare settings, as missing actual positive cases (such as high-risk individuals) can lead to serious repercussions.

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

Where:

- *FN*: False Negatives — high-risk cases incorrectly predicted as low-risk.

A high recall indicates that the model successfully identifies the majority of actual high-risk cases, minimizing missed detections.

L. F1-Score

Since precision and recall often involve a trade-off, the F1score is used to balance them by computing their harmonic mean. The F1-score provides a single, balanced metric that is particularly useful for imbalanced datasets, which are common in clinical data, where one class (e.g., high-risk cases) may be less frequent.

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

The F1-score ranges from 0 to 1, with higher values indicating better overall classification performance that balances both false positives and false negatives.

Alongside precision, recall, and F1-score, additional evaluation methods were utilized to gain a deeper insight into the model's performance. The confusion matrix served to

 ISBN: 978-9982-95-500-3

clearly contrast predicted and actual class labels, illustrating the counts of: Where:

True Positives (TP): High-risk cases correctly identified as high-risk.

True Negatives (TN): Low-risk cases correctly identified as low-risk.

False Positives (FP): Low-risk cases incorrectly predicted as high-risk (Type I error).

False Negatives (FN): High-risk cases incorrectly predicted as low-risk (Type II error).

Analyzing these values enabled the identification of specific misclassification types, revealing consistent error patterns and potential biases. The confusion matrix was generated using Matplotlib, providing a clear numerical view of predicted versus actual outcomes and supporting the interpretation of classification performance. This insight highlighted areas where additional model tuning or data preprocessing could further improve results.

In addition to the confusion matrix, the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) metric was employed to assess the model's ability to discriminate between high-risk and low-risk cases across varying classification thresholds. The ROC-AUC offers a single, thresholdindependent measure of performance: an AUC of 1.0 indicates perfect discrimination, a value of 0.5 corresponds to random guessing, and values closer to 1.0 represent stronger discriminative power. This metric is particularly useful in clinical applications, as it reflects the model's ability to maintain an appropriate balance between sensitivity and specificity under different operating conditions, offering a robust summary of its predictive capability without requiring graphical visualization.

M. Integration of Explainability

Once the model was trained, SHAP and LIME were employed to interpret the outputs of the DistilBERT model. The insights obtained from these methods were validated both quantitatively, through feature importance rankings, and qualitatively, through individual example explanations.

N. Deployment

The final product is a Flask-based web application that provides real-time sentiment predictions. Developed using an agile methodology, the system underwent iterative cycles of implementation, testing, and refinement. The user interface presents a text input area, real-time sentiment predictions, emotion probabilities, and mental health assessment results.

IV. RESULTS

This section presents the evaluation results of the proposed mental health detection framework. The analysis focuses on both quantitative performance metrics and qualitative interpretability, reflecting the dual emphasis of this study on accuracy and explainability. Two models were compared: a baseline Random Forest classifier trained on TF-IDF features and a fine-tuned DistilBERT transformer model. Their effectiveness was assessed using a balanced test dataset of social media posts, with performance measured through accuracy, precision, recall, F1score, and ROC-AUC. In addition to numerical evaluation, interpretability was examined using SHAP and LIME explanations, with validation by mental health professionals to determine clinical relevance. Given the ethical sensitivity of applying artificial intelligence to mental health, the chapter also discusses error analysis, computational efficiency, visualization outputs, and the risks associated with misclassification.

A. Dataset Summary for Evaluation

The evaluation was conducted on a held-out test set comprising 2,692 anonymised posts, which were drawn from the pre-processed dataset of 26,925 samples. Stratified sampling ensured that both high-risk and low-risk categories were equally represented, thereby preventing class imbalance from skewing the results. The dataset preserved the linguistic richness typical of social media communication, including colloquial expressions, abbreviations, and emotive language. Importantly, all personally identifiable information (PII) was removed during preprocessing to safeguard privacy. This test set provided a representative environment for evaluating the robustness and generalisability of the models.

B. Classification Performance

The comparison between DistilBERT and the Random Forest baseline revealed that the transformer model consistently outperformed the traditional approach across all evaluation metrics. DistilBERT achieved an accuracy of 92 percent, compared to 85 percent for the baseline. Precision improved from 82 to 89 percent, recall rose from 80 to 91 percent, and the F1-score increased from 81 to 90 percent. Furthermore, the ROC-AUC for DistilBERT was 0.94, significantly higher than the 0.88 achieved by the baseline. Statistical tests confirmed that these differences were significant, with McNemar's test indicating p-values of less than 0.05 across all major metrics. These results demonstrate that transformer-based architectures are more effective at capturing semantic nuance and contextual meaning in mental health-related text.

TABLE I
PERFORMANCE COMPARISON BETWEEN DISTILBERT AND TF-IDF + RF

Metric	DistilBERT	TF-IDF + RF	Difference	p-value (McNemar)
Accuracy	92%	85%	+7%	<0.01
Precision	89%	82%	+7%	<0.05
Recall	91%	80%	+11%	<0.01
F1-Score	90%	81%	+9%	<0.01
ROC-AUC	0.94	0.88	+0.06	<0.05

 ISBN: 978-9982-95-500-3

V. DISCUSSION

In this study, an accurate and precise mental health detection system was developed using social media data through advanced machine learning methods, complemented by explainable artificial intelligence (XAI). The model achieved high accuracy and precision. One major challenge encountered was the class imbalance in the dataset, where posts labeled as highrisk were notably fewer than those identified as neutral or lowrisk. This imbalance posed a risk of bias toward the majority class, potentially limiting the model's ability to detect critical high-risk cases. To address this issue, stratified sampling was applied during training, and threshold adjustments were implemented to improve sensitivity to minority cases, thereby enhancing overall prediction accuracy. Given the sensitive nature of mental health applications, a strong emphasis was placed on interpretability. Techniques such as SHAP and LIME were used to generate comprehensive explanations of the model's predictions, highlighting the key terms that influenced classifications. These approaches helped reduce the complexity often associated with deep learning models, making the results more understandable for non-expert users.

The DistilBERT transformer model outperformed the baseline TF-IDF combined with a Random Forest classifier across key performance metrics, including accuracy, precision, recall, and F1-score. This demonstrates the advantage of transformerbased contextual embeddings over traditional feature-based methods in capturing the semantic and contextual nuances of text. Nevertheless, balancing model complexity with clear, interpretable explanations remains an ongoing challenge. Ethical safeguards were integrated into the system, including confidence thresholds to minimize harmful errors and measures to ensure fair representation across demographic groups, thereby promoting equitable performance.

VI. CONCLUSION AND RECOMMENDATIONS

This study investigated the application of explainable artificial intelligence (XAI) techniques to enhance the detection of depression using transformer-based natural language processing (NLP) models. The fine-tuned DistilBERT model demonstrated strong predictive performance. Additionally, the integration of SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) provided crucial transparency into the model's decision-making. These explainability methods tackled challenges associated with the opaque nature of deep learning models in healthcare, thereby enhancing the system's trustworthiness and clinical relevance. The results highlight the importance of striking a balance between predictive accuracy and interpretability when developing AI tools for mental health assessment. Although the current system shows promising results, further work is needed to expand its capabilities. Future research could explore multi-label classification to detect a broader range of mental health conditions beyond depression. Additionally, incorporating real-time alert systems linked to support services may increase the system's practical effectiveness in clinical and community environments. Efforts should also focus on improving accessibility by developing mobile-based platforms to reach wider populations. The application of active learning could support ongoing model refinement with new, informative data. Migrating inference to scalable cloud platforms, such as the Hugging Face API, may facilitate efficient deployment and adaptability in real-world settings. Finally, evaluation of the system with end users is necessary to gather feedback, ensuring its usability, effectiveness, and alignment with user needs in practical applications.

REFERENCES

I. o. H. M. a. Evaluation, "Depressive disorder (depression)," World Health Organization, 2023.

A. P. O. P. W. Adewale Abayomi Adeniran, "Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical," World Journal of Advanced Research and Reviews, vol. 23, no. 3, p. 2647–2658, 2024.

A. k. ,. B. E. A. a. M. R. Bihi Sabiri, "Analyzing BERT's Performance Compared toTraditional Text Classification Models," in 25th International Conference on Enterprise Information Systems, 2023.

S. P. P. S. a. M. M. N. Patel, "XAIForCOVID-19: A comparative analysis of various explainable AI techniques for COVID-19 diagnosis using chest X-ray images," in International conference on computer vision and image processing, 2023.

M. M. F. a. Z. M. F. V. Pitroda, "An explainable AI model for interpretable lung disease classification," in IEEE International Conference on Internet of Things and Intelligence System, 2021.

P. Y. M. S. K. a. J. C. M. Bhandari, "Exploring the capabilities of a lightweight CNN model in accurately identifying renal abnormalities," Applied Sciences, vol. 13, no. 5, p. 3125, 2023.

K. M. G. a. L. L. L. J. H. Ong, "Comparative analysis of explainable artificial intelligence for COVID-19 diagnosis on CXR image," in International Conference on Image and Signal Processing and their Applications, 2021.

F. L.-P. L. M.-S. M. D.-M. a. A. C. J. Civit-Masot, "A lightweight xAI approach to cervical cancer classification," Medical & Biological Engineering & Computing, vol. 62, no. 8, p. 2281–2304, 2024.

P. I. N. S. D. M. a. P. Y. N. Gnanavel, "Interpretable cervical cell classification: A comparative analysis," in International Conference on Advances in Computing, 2024. in International Research Conference on Smart Computing and Systems Engineering

. G. P. I. D. M. a. P. Y. N. Sritharan, "EnsembleCAM: Unified visualization for explainable cervical cancer identification," in International Research Conference on Smart Computing and Systems Engineering, 2024.

W. H. Organization., "Mental health of older adults," 2023.

M. Zimmerman, "The value and limitations of self-administered questionnaires in clinical practice and epidemiological studies," World Psychiatry, vol. 23, no. 2, p. 210–212, 2024 .

M. S. P. G. T. J. V. K. V. a. A. H. Pradeep Kumar Tiwari, "A Study on Sentiment Analysis of Mental Illness Using Machine Learning Techniques," [Online]. Available: https://www.kaggle.com/code/sasakitetsuya/mental-health-classifier-nlp. [Accessed 13 March 2025].

A. I. M. T. Nasir Musa Imam1, "Explainable Artificial Intelligence (XAI) Techniques To Enhance Transparency In Deep Learning Models," IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 26, no. 6, pp. 29-36, Nov. – Dec. 2024.

T. D. A. K. P.-M. L. S. S. Antoine Hudon(B), "Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence," in Information Systems and Neuroscience, Springer Nature Switzerland AG, 2021, pp. 237-246.

. U. N. T. S. K. B.&. M. Z. A. Daniyal Alghazzawi, ”Explainable AIbased suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique,” Scientific Reports, 2025.

R. S. F. A. P. N. S. R. M. R. Tazrin Rahman1, ”Text-Based Data Analysis for Mental Health Using Explainable AI and Deep Learning,” in Networking and Parallel/Distributed Computing Systems , 2024, pp. 17-31.

S. Z. Y. Q. a. D. W. Elma Kerz1, ”Toward explainable AI (XAI) for mental health detection based on language behavior,” Frontiers in Psychiatry, 2023.

P. K. ,. G. S. Leon Kopitar, ”Hybrid visualization-based framework for depressive state detection and characterization of atypical patients,” Journal of Biomedical Informatics , vol. 147, no. 1, 2023.

B. S. Gulsum Alicioglu, ”A survey of visual analytics for Explainable Artificial Intelligence methods,” Computers & Graphics, vol. 102, pp. 502-520, 2022.

P. B. L. Christoph Tauchert, ”Crowdsourcing Data Science: A Qualitative Analysis of Organizations’ Usage of Kaggle Competitions,” in Hawaii International Conference on System Sciences, 2020.

 ISBN: 978-9982-95-500-3