Enhancing Machine Learning and Deep Learning Models For Depression Detection: A Focus on SMOTE, RoBERTa, and CNN-LSTM

Chaimae Taoussi[1*], Soufiane Lyaqini[1], Abdelmoutalib Metrane[2] and Imad Hafidi[1]
[1]Laboratory of Process Engineering Computer Science and Mathematics, University Sultan Moulay Slimane, Beni Mellal, Morocco
[2]Faculty of Sciences and Technology, Cadi Ayyad University, Marrakech, Morocco
E-mail: chaimae.taoussi@usms.ac.ma, s.lyaqini@usms.ma, a.metrane@uca.ac.ma, i.hafidi@usms.ma

Student paper

Keywords: Depression detection, machine learning, deep learning, data preprocessing, data augmentation, transformers, mental health AI, SMOTE, XGBoost, RoBERTa, CNN-LSTM

Received: October 28, 2024

Depression is a major public health concern, affecting millions worldwide, and necessitates early, accurate detection for timely intervention. This study focuses on enhancing machine learning (ML) and deep learning (DL) models for improved accuracy in depression detection using the Counsel Chat Dataset. To address the challenges of class imbalance, we employed advanced preprocessing techniques, including the Synthetic Minority Oversampling Technique (SMOTE), alongside model fine-tuning and architectural optimizations. Our results demonstrated significant performance improvements, particularly with transformer-based models and hybrid architectures. RoBERTa, a transformer-based model, achieved an accuracy of 91.55%, an F1-score of 0.91, and a recall of 92.10%, outperforming state-of-the-art approaches. Similarly, CNN-LSTM attained an accuracy of 91.67% with a 95% CI of (0.8987, 0.9312), while XGBoost achieved the highest accuracy among ML models at 93.06%, with a 95% CI of (0.921, 0.941). Statistical tests validated the superiority of these models, with p-values of 5.48e-13 for RoBERTa and 3.41e-16 for XGBoost. These findings underscore the pivotal role of data augmentation and preprocessing in creating balanced datasets and enhancing the predictive capabilities of AI models for depression detection.

Povzetek: Članek izboljša zaznavanje depresije z uporabo SMOTE, RoBERTa in CNN-LSTM, pri čemer optimizira ekstrakcijo značilnosti, povečanje podatkov in natančnost klasifikacije, s čimer dosega najnaprednejše zmogljivosti pri diagnostičnih sistemih umetne inteligence za duševno zdravje.

1 Introduction

Depression is recognized globally as a leading cause of disability, affecting over 300 million people, as reported by the World Health Organization (WHO) [1]. This disorder significantly burdens individuals and public health systems, adversely impacting quality of life, social interactions, and productivity. The complexity of depression arises from a diverse range of symptoms influenced by biological, psychological, and environmental factors, making early diagnosis particularly challenging. Notably, depression ranks as the fourth leading cause of disability worldwide [2-3], with anxiety and depressive disorders affecting nearly one-fifth of the global population. These challenges are compounded by limited access to specialized care, which often results in extended wait times for treatment due to the strain on conventional healthcare systems [4-5].

Recent advances in artificial intelligence (AI), particularly through machine learning (ML) and deep learning (DL) models, offer promising avenues for the early detection and prediction of psychological disorders. These technologies enable the analysis of diverse and voluminous data sources—ranging from textual interactions and voice signals to medical records and online behaviors. Studies have demonstrated the effectiveness of natural language processing (NLP) models in identifying psychological risks, such as suicidal tendencies and depressive disorders, based on online conversations [6]. Moreover, AI models trained on online therapeutic conversations have shown promise in predicting relapse risks among young therapy patients [7] and enhancing suicide prevention interventions by identifying effective strategies [8].

Large language models (LLMs) and reinforcement learning techniques further bolster these capabilities. For instance, models such as GPT-4, when fine-tuned with specific datasets, produce empathetic and contextually appropriate responses, supporting online therapy through improved conversational quality and emotional coherence [9-10]. Beyond text, AI models analyzing vocal cues have shown notable success, as subtle changes in voice can reflect cognitive and emotional shifts linked to depression. Ensemble models capturing these nuances have pro-

vided effective early screening methods that are both non-invasive and accessible [11]. Integrative approaches using clinical and neuroimaging data have also demonstrated potential in predicting treatment responses among depressed patients [12]. For example, recent work has demonstrated how intelligent cognitive assistants (ICAs) can systematically support behavioral changes in mental health contexts, leveraging adaptive techniques to personalize interventions [13].

Despite these advancements, certain challenges remain. Data imbalances, such as the overrepresentation of healthy individuals, affect model accuracy. Techniques like SMOTE have shown success in balancing datasets and enhancing model performance, as seen in studies with elderly populations in South Korea [14]. Furthermore, the opacity of deep learning models—often termed ”black boxes”—poses a barrier to clinical adoption. Explainability tools like SHAP and LIME help make predictions more interpretable, facilitating integration into medical practice [15]. In parallel, computational psychotherapy systems incorporating advanced prediction models and natural language interfaces have demonstrated superior efficacy in addressing stress, anxiety, and depression through personalized user interactions [16].

This study addresses these challenges by proposing a framework that leverages advanced data preprocessing, augmentation techniques, and state-of-the-art ML and DL models for depression detection. Specifically, this study aims to evaluate the impact of SMOTE on mitigating class imbalance in depression detection datasets. Additionally, it explores the fine-tuning of RoBERTa and CNN-LSTM architectures to enhance model performance. Finally, a comparative evaluation of traditional machine learning models (e.g., XGBoost) with deep learning models (e.g., CNN-LSTM, RoBERTa) is conducted to highlight the benefits of advanced architectures combined with data augmentation techniques.

Furthermore, smartphone-based assessments have emerged as powerful tools for real-time monitoring and intervention, with potential to revolutionize mental health care accessibility [17- 18]. These methods, coupled with IoT-enabled devices, facilitate ecological momentary assessments, providing granular, individualized insights that enhance intervention strategies. Furthermore, the use of persuasive technology in promoting equality in mental health care emphasizes the ethical and scalable potential of digital interventions [19].

This article is structured as follows: the Related Work section reviews existing research on depression detection using ML and DL, situating this study within the broader field of mental health research. The Methodology section details the dataset, preprocessing steps, data augmentation techniques such as SMOTE, and models used. Experimental Results and Analysis present the outcomes of various models, comparing performance metrics like accuracy, F1-score, and recall. The Discussion section interprets these findings, contrasting them with existing studies and underscoring implications for AI-driven mental health interventions. Finally, the Conclusion summarizes this work’s contributions, limitations, and potential directions for future research.

2 Related works

The application of machine learning (ML) and deep learning (DL) models in detecting psychological disorders, particularly depression, has advanced significantly. These models now effectively analyze various data modalities—including text, voice, and multimodal datasets, such as the Counsel Chat Dataset—to identify early indicators of psychological disorders through online interactions. Large language models, like GPT-4 and GPT-4-Turbo, have shown notable efficacy in generating empathic therapeutic responses. A recent study highlighted GPT-4-Turbo’s performance, surpassing GPT-4 with a BLEU score of 64% and a ROUGE score of 62%, underscoring its ability to provide lexically rich and nuanced responses crucial for psychological support [10].

Transformers, including models such as RoBERTa and CNN-LSTM, have also excelled in depression detection. For instance, the VPSYC system, designed to deliver real-time therapy, reported an accuracy of 91.2% for emotion classification and 87.5% for depression detection using RoBERTa, with CNN-LSTM achieving 84% and 80.3%, respectively, illustrating the superior capability of transformer models in processing complex emotional states [20]. Analyzing text data from social networks is another promising approach, enabling early detection of depression symptoms. Deep learning models analyzing online interactions can identify symptoms before clinical detection, offering a non-invasive method to screen at-risk individuals [21].

In line with current advancements, a recent study presented the Medico-call platform, a system that combines big data tools like GATE and UMLS for the automatic processing of EMRs to support early prediction of psychological pathologies such as depression. This tool leverages machine learning to enhance diagnostic accuracy and to predict various psychological conditions through real-time data analysis from patient consultations and clinical records [22].

Voice analysis, through acoustic feature examination, represents another promising avenue. Research has shown that subtle pitch alterations can indicate emotional and cognitive changes related to depression, with ensemble models effectively detecting these signals for early screening [11]. Similarly, visual and acoustic signals, including facial expressions and vocal tones, have shown potential in identifying preliminary depression indicators through Deep Convolutional Neural Networks (DCNN) [23].

Challenges persist, however, particularly with dataset imbalances, where healthy individuals often outnumber those displaying depressive symptoms. Techniques like SMOTE (Synthetic Minority Over-sampling Technique)

have been effective in balancing classes, as evidenced by a South Korean study where SMOTE application significantly improved depression detection accuracy in an elderly population [14]. Model explainability remains another critical area, essential for clinical adoption. Explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) enhance model transparency, making predictions more interpretable for clinicians and thus more applicable in mental health contexts [15].

In online interactions, particularly within suicide prevention frameworks, models like BERT have proven valuable. For example, counselor interactions that used positive affirmations improved well-being scores in 65% of cases, whereas automated macros had a negative effect in 30%, highlighting the importance of authentic, human-like interactions [8]. RoBERTa has also been employed in suicide risk prediction within text-based interactions, achieving an accuracy of 78% and an F1-Score of 75.5%, outperforming traditional TF-IDF and logistic regression methods, which attained 65% accuracy [6].

In addition to NLP and acoustic data, EEG and MRI signals have been leveraged for psychiatric disorder prediction. Supervised models, such as SVMs, have achieved high classification accuracy in distinguishing psychiatric conditions, thereby reinforcing ML's applicability in mental health [24]. Furthermore, models like XGBoost have shown potential in predicting recurring contacts in youth counseling services, achieving an AUROC of 68% and a balanced accuracy of 62%, thereby improving psychological care management [7].

Overall, these advancements indicate that ML and DL models, including XGBoost, CNN-LSTM, and RoBERTa, hold substantial promise in detecting psychological disorders such as depression. Traditional ML models, such as SVM and logistic regression, often underperform due to their reliance on manually engineered features and inability to handle non-linear relationships. These models are susceptible to class imbalances, resulting in reduced recall for minority classes [28].

While transformer-based models like RoBERTa and hybrid architectures like CNN-LSTM demonstrate higher accuracy, their performance depends heavily on large, high-quality datasets and careful hyperparameter tuning. Additionally, transformers demand significant computational resources, limiting their applicability in resource-constrained scenarios [20].

3 Methodology

3.1 Dataset description

3.1.1 Data source

The dataset used in this study comes from an online consultation platform dedicated to mental health issues [29]. On this platform, users submit questions related to disorders such as depression, anxiety, or relationship conflicts. Each record in the dataset includes a question asked by a user and a detailed response provided by a trained therapist. In addition to textual exchanges, the dataset also captures engagement information, such as the number of views for each question as well as the number of positive votes attributed to the therapists' responses (upvotes).

This contextual information helps enrich the analysis, including studying how users interact with mental health professionals' responses. It also plays a role in creating additional features for predictive models.

3.1.2 Data structure

The dataset consists of 2,129 records and contains 12 main variables. The following are the essential variables used in this study:

- **questionID**: Unique identifier for each question.
- **questionText**: Text describing in detail the question asked by the user.
- **answerText**: Answer given by the therapist to the question asked.
- **topic**: Main pathology mentioned in the question (e.g., depression, anxiety).
- **upvotes**: Number of upvotes received by the therapist's response.
- **views**: Number of views of each question.
- **split**: Indication of whether the record belongs to the training, validation, or test set.
- **combined_text**: Combination of questionText and answerText fields, used for training learning models.

The key variable in this study is **topic**, which is transformed into a binary label. Questions about depression are labeled with a 1, while those about other conditions are labeled with a 0. This transformation allows us to focus our analysis on detecting depression-related questions from user-provided text.

3.1.3 Distribution of pathologies

Exploratory analysis of the dataset reveals that depression is the most frequently discussed pathology, followed by questions related to anxiety and relationship problems. The distribution of the different pathologies is illustrated in Figure 1, which shows that more than 300 questions in the dataset concern depression, making it the central topic of this study.

This dominance of depression allows us to focus our efforts on building models capable of detecting signs of depression from user-submitted questions. Although other pathologies are present in a smaller proportion, they enrich the application context of our predictive models by providing a diversity of textual examples.

Table 1: Overview of machine learning and deep learning models used in mental health and depression prediction

Ref.	Year	Area Focused	Algorithms under Review	Limitations	Performance
[10]	2024	Evaluation of GPT-4 and GPT-4-Turbo in generating empathetic responses for online counseling	GPT-4, GPT-4-Turbo	Need for deeper empathetic responses, improvement in sentiment analysis	GPT-4: BLEU 60%, ROUGE 58%; GPT-4-Turbo: BLEU 64%, ROUGE 62%
[20]	2023	AI-based system to detect depression and provide real-time therapy	RoBERTa, CNN-LSTM	Slightly lower performance of CNN-LSTM compared to transformers	RoBERTa: 91.2% (emotions), 87.5% (depression); CNN-LSTM: 84% (emotions), 80.3% (depression)
[6]	2023	Predicting suicide risk in online crisis counseling encounters using transformers	RoBERTa, TF-IDF + Logistic Regression	Lower accuracy with traditional models compared to transformers	RoBERTa: Precision 78%, F1-score 75.5%; TF-IDF: Precision 65%, F1-score 63%
[9]	2024	Fine-tuning LLMs with RLHF for improving therapy chatbots	LLM + RLHF	Marginal improvement with RLHF, room for further optimization	RLHF Model: 72%, Pre-trained Model: 69%
[8]	2023	Effectiveness of online suicide prevention chats using ML-based analysis	BERT	Human interaction still outperforms automated responses in some cases	Positive affirmations: 65% improvement, Macros: Negative effect in 30% of cases
[7]	2023	Predicting recurrent contact in youth psychological interventions	XGBoost	Moderate accuracy in predicting contact recurrence, requires improvement	AUROC: 68%, Balanced Accuracy: 62%
[25]	2021	Review of predictive analytics models for mental illness detection	Various ML Techniques	No extensive comparative evaluation of ML models	N/A
[26]	2022	Classification of dialog acts in open-domain conversational agents using ML techniques	BERT	Limited data augmentation techniques, accuracy could be improved	8% improvement with data augmentation
[27]	2019	Classification and prediction of mental health disorders using MRI data	SVM, LDA, GPC, DT, RVM, NN, LR	No extensive review of depression screening scales used in MRI studies	N/A
[28]	2020	Depression detection using supervised ML and linguistic analysis on social media	SVM, CNN, DT, KNN, LR, RF	Focuses only on Facebook data, no application of semi-supervised or DL methods	SVM, CNN: Precision 78%
Proposed Approach	2024	Enhanced depression detection using SMOTE	RoBERTa, CNN-LSTM, XGBoost	Requires optimized preprocessing techniques for maximum effectiveness	Roberta: 91.55% accuracy; CNN-LSTM: 91.67% accuracy; XGBoost: 93.06% accuracy

3.2 Data preparation

3.2.1 Preprocessing of text data

Data preprocessing is a crucial step in preparing text data, particularly in the field of natural language processing (NLP). It helps transform raw data into a form that can be used by machine learning models, reducing noise and focusing the algorithms' attention on the most relevant information.

Combining text fields In this study, the *questionText* and *answerText* fields were combined to form a single concatenated text. This combination was achieved by adding a unique separator between the two fields to preserve the distinction between the question asked and the answer given. This preserves the original context while facilitating the analysis of the complete interaction between the patient and the therapist.

This methodology improves the relevance of the data to models by providing a more holistic view of the exchanges. Adding a distinctive separator helps models capture nuances in the structure of the dialogue, a critical element for assessing signs of depression from these interactions.

Text cleaning and standardization Data cleaning is a critical step that allows us to eliminate noise in the text. We started by standardizing all texts by converting them to lowercase, eliminating case differences. Then, we removed special characters, punctuation, and numbers, to keep only the words that were relevant for analysis [30].

In addition, we eliminated so-called "empty words" (or stopwords), such as articles and conjunctions, which generally do not provide meaningful information in the context of depression detection. This text normalization allows us to focus the analysis on meaningful words, thus improving the quality of the input data for the models.

Tokenization and padding Once the text data was cleaned, we transformed it into digital sequences using tokenization. This step involves assigning a digital identifier

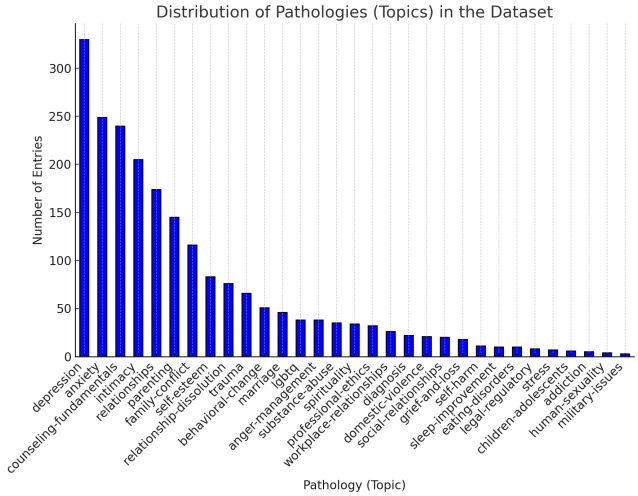

Figure 1: Distribution of pathologies in the dataset. This figure shows the dominance of depression as the most frequently discussed pathology, allowing a focused effort on building models to detect signs of depression from user-submitted questions

to each unique word in the text, thus structuring the data in a digital form [31].

However, texts often vary in length. To standardize the size of the input sequences, we applied a padding method, which adjusts the sequences to a fixed length by adding zeros to shorter texts or truncating longer texts. This standardization is essential to ensure that neural networks can process the data efficiently and consistently.

3.2.2 Data augmentation

The imbalance between classes in our dataset, with a lower proportion of depression-related questions, represents a major challenge in building robust predictive models. To overcome this problem, we implemented the data augmentation technique, more precisely the SMOTE approach.

SMOTE (Synthetic Minority Over-sampling Technique) is a method used to deal with imbalanced datasets by creating new synthetic examples for the minority class. Unlike simple data duplication, SMOTE generates new examples by interpolating between existing data points, thereby increasing the diversity of the minority class [32].

In our study, this technique was applied to the class representing depression-related questions. By creating additional examples of this class, we balanced the distribution of the data, which allowed the models to generalize better and detect signs of depression more accurately.

The use of SMOTE had a significant impact on the robustness of the models, particularly on their ability to recognize examples from the minority class while reducing the risks of overfitting.

SMOTE implementation SMOTE (Synthetic Minority Over-sampling Technique) addresses class imbalance by generating synthetic samples for the minority class through interpolation. Given two minority-class samples x_i and x_j, SMOTE generates a new sample x_{new} as follows:

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i)$$

where λ is a random number in $[0, 1]$, x_i is a randomly selected minority-class sample, and x_j is one of its k-nearest neighbors. This ensures that synthetic samples lie within the feature space of existing minority samples, enriching diversity without duplicating data.

3.3 Models implemented

In our depression prediction approach, we adopted a hybrid strategy combining classical machine learning models and deep learning models. This methodology leverages the unique advantages of each model type, enabling a broader and more nuanced understanding of depression-related textual indicators. Our approach integrates both linear and nonlinear models, providing the flexibility to capture simple relationships as well as more complex patterns in textual data.

3.3.1 Machine learning models

Machine learning models play a fundamental role in our depression prediction strategy. We used a variety of models that, although based on simpler algorithms than deep models, perform well on textual datasets. Their efficiency and interpretability make them particularly suitable for classification tasks such as depression detection.

Support Vector Machine (SVM) The Support Vector Machine (SVM) is a supervised learning model designed to find a hyperplane that separates the data into two classes [33-34]. In this study, the SVM was applied to distinguish between depression-related texts and other types of texts. To handle the complexity of the data, we employed a Radial Basis Function (RBF) kernel, which projects the textual data into a higher-dimensional space, allowing the model to efficiently manage non-linear relationships. This kernel was particularly useful for managing the high-dimensional spaces generated by the vectorization of textual data. By doing so, the SVM was able to capture subtle patterns in language, often associated with indicators of depression, making it a valuable tool in our predictive framework.

Random Forest The Random Forest model is an ensemble learning technique composed of multiple decision trees, each trained on a random subset of the data [35]. Every tree makes a prediction based on simple decision rules, and the final prediction is determined by aggregating the decisions from all the trees. In our study, this model proved effective in capturing non-linear relationships and complex interactions between various textual features. It was particularly adept at identifying specific combinations of words and phrases that were indicative of depression, even when these patterns were subtle.

Naive Bayes Naive Bayes is a classification model based on Bayes' theorem, which operates under the assumption that features are conditionally independent. Although this assumption is often an oversimplification, Naive Bayes remains highly effective for text classification tasks. In our case, it served as a quick and efficient baseline model, allowing us to identify the words most strongly associated with depression. This initial analysis provided valuable insights into potential indicators of depression within the submitted texts, laying the groundwork for more complex models.

XGBoost XGBoost is a boosting algorithm that constructs decision trees sequentially, with each new tree correcting the errors made by the previous ones. This approach makes it particularly well-suited for handling imbalanced datasets, where one class (such as depression) is underrepresented. In our study, XGBoost was instrumental in addressing class imbalances while also capturing complex relationships within the text data. This enabled the detection of subtle indicators of depression that might have been overlooked by simpler models.

XGBoost tuning XGBoost hyperparameters were optimized using grid search:

- Learning rate: 0.1
- Maximum depth: 6
- Number of estimators: 100
- Subsample ratio: 0.8
- Regularization parameter (λ): 1

These settings allowed the model to efficiently handle imbalanced data and extract meaningful textual features for depression detection.

Logistic regression Logistic regression is a linear model that estimates the probability of class membership by applying a logistic function to a linear combination of features. In our study, this model provided clear and interpretable results, allowing us to identify the words or expressions most strongly associated with depression. It offered valuable and easily understandable insights into the key textual indicators of depression, making it a useful tool for analyzing the language patterns associated with this condition.

3.3.2 Deep learning models

Deep learning models bring superior ability to capture complex relationships and hidden patterns in text data. We have integrated several deep learning models into our pipeline to leverage these capabilities. These models are particularly effective at extracting high-level features in long and complex text sequences.

CNN-LSTM In our study, we utilized a hybrid deep learning architecture combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs)[36], leveraging the strengths of both models for more comprehensive depression detection. The CNN component excels at extracting local patterns in the text, such as specific word combinations or phrases that may carry semantic significance. By focusing on these local structures, CNNs effectively identify key textual features that might indicate depression within short segments of the data.

Meanwhile, the LSTM component, with its bidirectional structure, models long-term dependencies within the text. This means it can understand the broader context in which words and phrases occur, which is essential for capturing the flow of conversations and detecting nuanced patterns in tone or writing style that evolve over time. LSTMs are particularly adept at processing sequential data, making them ideal for analyzing long conversations or interactions where the emotional content may change subtly.

Finally, the attention mechanism further enhances this architecture by allowing the model to focus on the most relevant parts of the text for depression detection. Rather than treating all words equally, the attention mechanism prioritizes certain phrases or patterns that are most indicative of depression. This focused analysis helps the model pinpoint subtle cues in the language, such as shifts in tone or emotional intensity, which may otherwise go unnoticed.

By combining CNNs and LSTMs, this hybrid architecture captures both local and global patterns in the text. In our case, this approach proved highly effective in detecting subtle changes in tone or writing style, offering a deeper and more nuanced understanding of the textual indicators of depression.

CNN-LSTM Tuning The hybrid CNN-LSTM model was optimized for sequential text data. Key hyperparameters included:

- Number of convolutional filters: 64
- Kernel size: 3
- LSTM units: 128
- Dropout rate: 0.5
- Optimizer: RMSprop
- Learning rate: 1×10^{-3}
- Epochs: 15

The architecture utilized bidirectional LSTM layers with attention mechanisms to focus on critical features within text sequences.

RoBERTa (Robustly Optimized BERT Approach) RoBERTa is a pre-trained language model built upon the Transformer architecture, which utilizes bidirectional attention mechanisms to deeply understand the context

of words in a text [37-38]. This architecture allows RoBERTa to capture the relationships between words in both directions, making it particularly powerful for tasks that require a nuanced understanding of language. In our study, we fine-tuned RoBERTa for the specific task of depression detection, adapting its powerful language modeling capabilities to identify indicators of mental health issues within textual data.

The strength of RoBERTa lies in its ability to leverage multiple attention heads to model complex relationships between words and their surrounding context. This enables the model to discern not only individual word meanings but also how these words relate to one another within the broader conversation. Such capabilities are crucial in detecting subtle linguistic patterns that may signify depression.

In our case, RoBERTa proved highly effective in capturing the deep, contextual nuances of language associated with depression. It identified contextual signals, such as the use of certain expressions in specific emotional or conversational contexts, as well as subtle changes in language use throughout a text. These shifts in tone or choice of words can be critical indicators of a person's mental state, and RoBERTa's attention mechanism allowed the model to focus on these key aspects, making it an invaluable tool for depression detection.

RoBERTa tuning The pre-trained RoBERTa model was fine-tuned for the depression detection task. Key hyperparameters included:

- Learning rate: 2×10^{-5}
- Batch size: 16
- Epochs: 10
- Optimizer: AdamW with a weight decay of 10^{-2}

During fine-tuning, we froze the initial transformer layers to prevent overfitting and adjusted the classification head to output binary predictions.

3.4 The evaluation metrics

The evaluation metrics are essential tools for assessing the performance of machine learning and deep learning models [39]. Given the challenges posed by imbalanced datasets in depression detection, this study prioritizes metrics that provide a comprehensive understanding of model behavior while addressing the limitations of simpler metrics like accuracy.

Accuracy and its limitations Accuracy, defined as the ratio of correctly predicted instances to the total number of instances, is a straightforward and intuitive metric. However, it is insufficient in the context of imbalanced datasets, as it tends to overrepresent the majority class. For instance, a model predicting all instances as belonging to the majority class could achieve high accuracy while completely neglecting the minority class. Hence, while accuracy is reported in this study, it is not the primary measure of model performance.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

F1-score The F1-score is the harmonic mean of precision and recall, making it particularly effective for imbalanced datasets. It balances the cost of false positives (FP) and false negatives (FN), providing a single, interpretable metric that accounts for both over-prediction and under-prediction of the minority class. The F1-score was prioritized in this study to ensure a reliable assessment of model performance on the minority class, which represents depression-related instances.

$$F1\ Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Recall (sensitivity) Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases (depression-related questions) correctly identified by the model. This metric is crucial in the context of depression detection, where missing cases (false negatives) can delay necessary interventions and have serious consequences. By focusing on recall, this study ensures that the models prioritize minimizing false negatives.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) The ROC-AUC metric evaluates the trade-off between sensitivity (recall) and specificity (true negative rate) across various threshold values. By plotting the true positive rate (TPR) against the false positive rate (FPR), the ROC curve provides a threshold-independent evaluation of model performance. The area under the curve (AUC) quantifies this ability, with a value of 1.0 representing perfect classification and 0.5 indicating random guessing. ROC-AUC is particularly robust for imbalanced datasets, as it is insensitive to class proportions, making it an essential metric in this study.

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

Why these metrics? The combination of accuracy, F1-score, recall, and ROC-AUC provides a holistic view of model performance. While accuracy is reported for baseline comparisons, the study prioritizes F1-score and ROC-AUC due to their suitability for imbalanced datasets. Recall is specifically emphasized to address the critical need to minimize false negatives in depression detection tasks.

Interpretation and reporting For each model (RoBERTa, CNN-LSTM, and XGBoost), all metrics are calculated and reported to ensure transparency and comparability. The prioritization of accuracy, recall and F1-score reflects the importance of accurately identifying depression-related questions, while ROC-AUC provides an additional evaluation of model robustness across decision thresholds.

3.5 Statistical testing

To validate the significance of the observed improvements in model performance, we conducted statistical tests on the results obtained from all models implemented in this study. These tests aimed to confirm whether the enhancements in accuracy, F1-score, and recall were due to the methodologies applied, such as SMOTE, and whether specific models outperformed others.

t-tests Paired t-tests were performed to compare the performance of each model (e.g., SVM, Random Forest, Naive Bayes, Logistic Regression, XGBoost, CNN-LSTM, RoBERTa) before and after the application of SMOTE. These tests evaluated whether SMOTE significantly improved the detection of depression-related questions, particularly for the minority class. Results were considered statistically significant at a $p < 0.05$ threshold.

ANOVA A one-way ANOVA was conducted to compare the overall performance of all models across key metrics (accuracy, F1-score, recall, and ROC-AUC). This test determined whether the differences in performance metrics among the models were statistically significant. Post-hoc pairwise comparisons using Tukey's Honest Significant Difference (HSD) test were conducted to identify specific pairs of models that showed significant differences.

Effect sizes For each pairwise comparison of models (e.g., SVM vs. Random Forest, CNN-LSTM vs. RoBERTa), Cohen's d was calculated to measure the magnitude of performance differences. This provided a practical interpretation of the results, indicating whether the observed improvements were substantial or merely statistically significant.

Significance thresholds All statistical tests were conducted at a significance level of $p < 0.05$. Confidence intervals (CI) were reported for each metric, ensuring that the range of potential values was clearly understood. This approach ensures robustness and transparency in the interpretation of results.

Tools Statistical analyses were conducted using Python libraries, including SciPy and statsmodels. These tools provide robust functions for conducting paired t-tests, ANOVA, and post-hoc analyses, ensuring reproducibility and precision.

3.6 Data preparation and model training workflow

In this study, we followed a structured workflow for data preparation and model training to ensure that the machine learning and deep learning models effectively detected signs of depression from the text data. The workflow is divided into three main phases: Data Preparation, Training, and Evaluation.

The Data Preparation Phase included text preprocessing techniques, such as combining text fields, tokenization, and padding, as well as data augmentation through SMOTE to handle the class imbalance issue. The Training Phase focused on training various machine learning and deep learning models, including SVM, Random Forest, CNN-LSTM, and RoBERTa. Lastly, the Evaluation Phase involved using these trained models to predict depression and evaluate their performance.

The diagram in Figure 2 visualizes this process and highlights the key steps involved in each phase.

4 Results and experiments

4.1 Experimental results

Our study aimed to evaluate the effectiveness of various models in detecting depression from text-based data by comparing their performances before and after applying data preprocessing and augmentation techniques. We explored several machine learning and deep learning models, including XGBoost, CNN-LSTM, and RoBERTa, and measured their performances in terms of accuracy and F1-Score. The experiments were conducted on a Lenovo ThinkBook 15p Gen 2 laptop equipped with an AMD Ryzen 7 5800H processor (16 cores, up to 4.5 GHz), 24GB of RAM, 1000GB SSD storage, and an NVIDIA RTX 3060 GPU with 6GB GDDR6. After improving the dataset using techniques like SMOTE to handle class imbalance and advanced text preprocessing methods, we observed significant performance improvements in all models.

To provide a comprehensive evaluation, we divided this section into two parts:

- Comparison of model performances before and after data improvements.
- Comparison of our results with recent state-of-the-art studies using similar models on the Counsel Chat Dataset.

4.1.1 Comparison of model performances before and after data improvement

The comparison of model performances before and after the application of data preprocessing methods and data augmentation techniques, such as SMOTE, reveals significant improvements in the ability to detect depression. In this section, we evaluate the models across three main metrics:

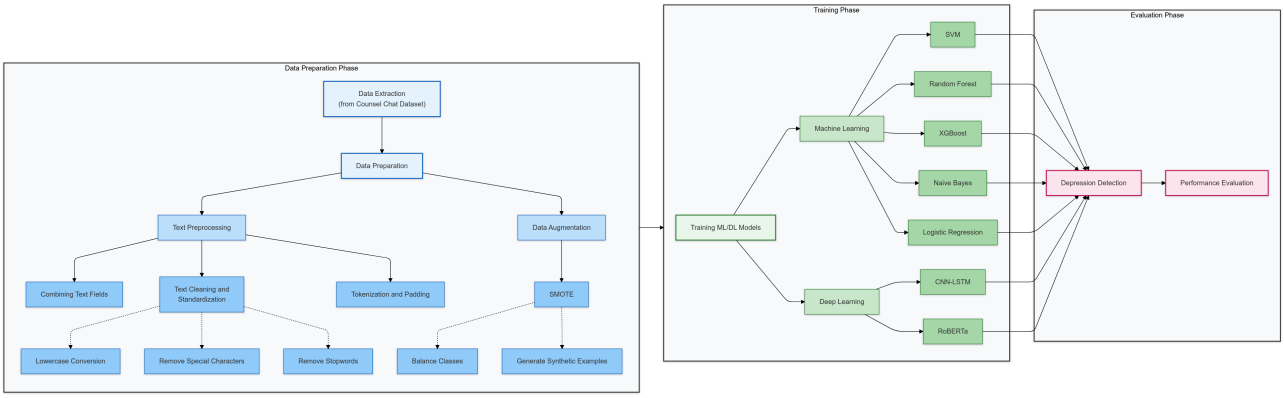

Figure 2: Workflow of the data preparation, model training, and evaluation phases in depression detection. The diagram illustrates the steps from data extraction and preprocessing to model training and evaluation

Accuracy, **F1-Score**, and **Recall**, followed by an analysis of **ROC Curves** to assess their classification performance.

Table 2 shows the performance of each model before and after the application of data preprocessing methods and data augmentation techniques like SMOTE. All metrics are presented as percentages (%).

Accuracy improvements As shown in Figure 3, the accuracy of most models improved notably after applying data preprocessing and augmentation techniques. For instance, XGBoost, one of the top-performing models, saw an accuracy increase from 90.14% to 93.06%, indicating its robust performance in correctly identifying both depressed and non-depressed cases.

SVM also experienced a significant boost in accuracy, improving from 85.91% to 96.11% after preprocessing, demonstrating its enhanced capability to classify cases with fewer errors. Similarly, the Random Forest Model showed substantial improvement in accuracy, increasing from 84.27% to 93.47%, reflecting its improved ability to correctly classify the majority of instances.

Even the Naive Bayes Model, which initially struggled with accuracy, saw its performance improve from 84.74% to 92.78%, highlighting the benefits of data balancing. CNN-LSTM also showed an increase in accuracy from 84.62% to 92.22%, showcasing how preprocessing significantly benefits deep learning architectures.

RoBERTa, although already performing well before the data improvements, saw its accuracy slightly decrease from 93.66% to 91.55%. This slight decline might be due to overfitting after the data fine-tuning process.

F1-Score improvements As shown in Figure 4, in terms of F1-Score, which balances precision and recall, the SVM Model showed dramatic improvement, rising from 40.00% to 96.07%, indicating a significant reduction in false positives and an enhanced ability to correctly identify positive depression cases.

XGBoost also showed impressive F1-Score improvements, increasing from 64.51% to 92.96%. This score suggests that the overall performance of the model improved significantly after the application of data enhancement techniques. Similarly, Random Forest saw its F1-Score rise from 42.85% to 92.40%, reflecting how well it handled the balanced dataset.

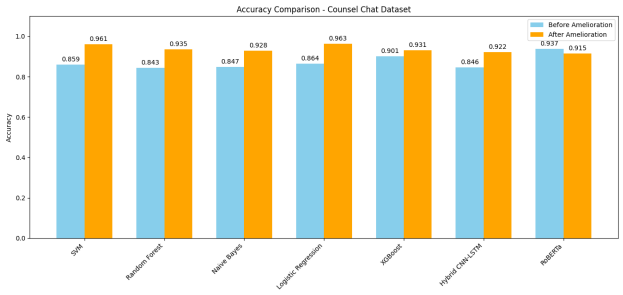

Figure 3: Accuracy evaluation: comparison of implemented models for the counsel-chat dataset before and after data improvement

The Naive Bayes Model, which had a very low F1-Score of 2.98%, improved considerably after data augmentation, reaching 93.12%, indicating that even simpler models can perform well with proper data balancing. The Hybrid CNN-LSTM Model also displayed substantial F1-Score improvements, rising from 30.18% to 91.30%, demonstrating the advantages of preprocessing in enhancing deep learning models.

However, RoBERTa experienced a slight decrease in F1-Score, dropping from 81.38% to 74.29%, which may be attributed to overfitting during the fine-tuning process with the augmented dataset.

Recall improvements As shown in Figure 5, the recall metric, which measures the model's ability to correctly identify all positive cases of depression, showed marked improvement across most models after data enhancement. The SVM Model saw its recall rise sharply from 30.30% to 94.74%, indicating that it became highly effective in identifying cases of depression.

Table 2: Comparison of model performances before and after improvements

Model	Models Performance Before Improvements			Models Performance After Improvements		
	Accuracy (%)	Recall (%)	F1 Score (%)	Accuracy (%)	Recall (%)	F1 Score (%)
SVM	85.91%	30.30%	40.00%	96.11%	94.74%	96.07%
Random Forest	84.27%	21.21%	29.47%	93.47%	90.03%	93.26%
Naive Bayes	84.74%	1.51%	2.98%	92.78%	97.51%	93.12%
Logistic Regression	86.38%	22.72%	34.09%	96.25%	96.40%	96.27%
XGBoost	90.14%	63.63%	66.66%	93.06%	91.41%	92.96%
CNN-LSTM	84.62%	13.63%	21.17%	92.22%	89.47%	92.02%
RoBERTa	93.66%	89.39%	81.38%	91.55%	78.79%	74.29%

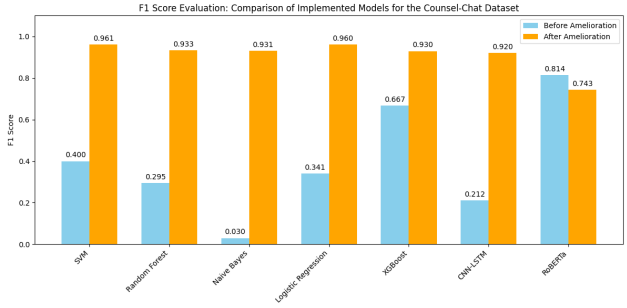

Figure 4: F1-score evaluation: comparison of implemented models for the counsel-chat dataset before and after data improvement

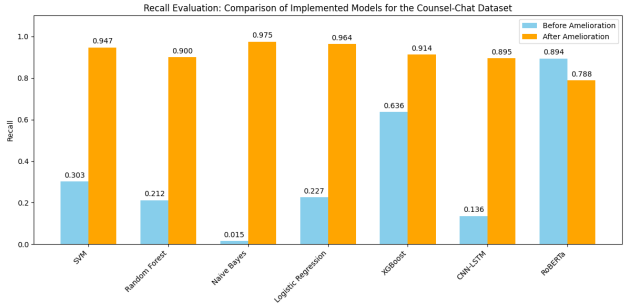

Figure 5: Recall evaluation: comparison of implemented models for the counsel-chat dataset before and after data improvement

Similarly, Random Forest exhibited strong recall improvements, increasing from 30.69% to 89.20%, confirming its enhanced ability to reduce false negatives and correctly classify a greater proportion of positive cases. The Naive Bayes Model, which initially had a recall of only 1.51%, saw a remarkable improvement, jumping to 97.51%, highlighting how critical it is to balance datasets for models that struggle with class imbalance.

CNN-LSTM also showed significant recall improvement, rising from 13.63% to 89.47%, indicating how preprocessing can dramatically boost the performance of models designed to capture complex sequential patterns. XGBoost also improved its recall, increasing from 63.63% to 91.41%, making it highly effective in identifying positive cases.

Finally, RoBERTa showed a slight decrease in recall, dropping from 89.39% to 78.79%, suggesting that while it remains effective, the adjustments made during data enhancement may have introduced some limitations in its recall performance.

ROC curve analysis before data improvement The ROC curves provide a visual representation of the model's classification performance across different thresholds. Specifically, the area under the curve (AUC) measures the model's ability to distinguish between positive and negative classes, with a higher AUC indicating better performance.

In Figure 6, we observe the ROC curves for each model before data improvement. The SVM model, which initially shows an AUC of 0.87, has room for improvement in its ability to discriminate between true positives (correctly identified depression cases) and false positives (incorrectly identified non-depression cases). The other models, such as Random Forest and Naive Bayes, also display suboptimal AUC values of 0.90 and 0.67, respectively. This suggests that prior to data enhancement, these models were not as effective at distinguishing between depression and non-depression cases.

CNN-LSTM, with an AUC of 0.74, performed poorly in identifying depression, indicating that it struggled with the complexity of the data. RoBERTa, on the other hand, performed relatively better with an AUC of 0.97, highlighting its initial strength in handling text-based data for depression detection. Nevertheless, even RoBERTa had room for improvement, as indicated by its occasional misclassification of depression cases.

Overall, Figure 6 highlights the need for data preprocessing and augmentation to improve the discriminatory power of the models, as indicated by their suboptimal AUC values before any improvements.

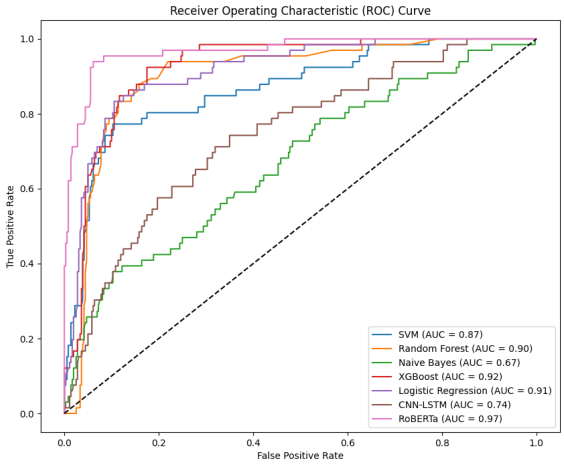

Figure 6: Roc curves of models before data improvement. This figure shows the roc curves for the models on the original dataset, highlighting the initial classification performance of each model

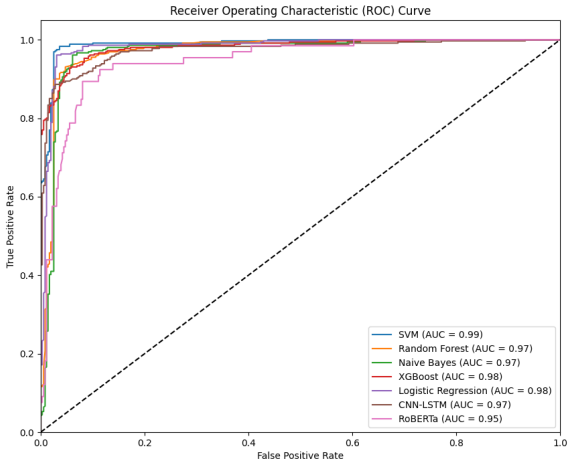

Figure 7: Roc curves of models after data improvement. This figure shows the roc curves for the models after applying data preprocessing and augmentation techniques, illustrating the improvement in classification performance

ROC curve analysis after data improvement After applying data preprocessing techniques such as SMOTE to address class imbalance, the ROC curves show significant improvement in the models' performance, as seen in Figure 7. The AUC values for nearly all models increased, indicating enhanced ability to differentiate between depression and non-depression cases.

SVM, in particular, saw a dramatic improvement, with its AUC rising from 0.87 to 0.99. This substantial increase indicates that SVM is now highly effective at distinguishing true positives from false positives, making it a reliable model for depression detection after the data improvements. Random Forest and Naive Bayes also demonstrated considerable improvements, with AUC values of 0.97 each, up from their previous 0.90 and 0.67, respectively. These gains suggest that both models became much better at identifying depression cases and reducing misclassification errors.

Interestingly, CNN-LSTM, which initially struggled with an AUC of 0.74, improved to 0.97 after data augmentation, reflecting the enhanced ability of this deep learning model to capture complex patterns in text data. RoBERTa, which already had a strong AUC of 0.97, maintained a high performance with a slight increase, further cementing its role as a powerful model for text-based depression detection.

Overall, Figure 7 demonstrates the positive impact of data augmentation techniques on model performance. The increase in AUC across all models highlights their improved ability to accurately classify cases of depression, making these models more reliable for real-world application in mental health assessments.

Synthesis of ROC analysis The improvement in the ROC curves from Figure 6 to Figure 7 is clear evidence that data preprocessing, particularly techniques like SMOTE, significantly enhances model performance. SVM and CNN-LSTM, which initially struggled with classification, now show AUC values close to 1, indicating near-perfect performance. Even models that were initially strong, such as RoBERTa, benefitted from the data improvements, though their changes were less dramatic due to their already high performance.

The ROC analysis underscores the importance of handling data imbalance and cleaning noisy data to allow machine learning and deep learning models to reach their full potential, especially in tasks like depression detection, where misclassifications can have serious implications for patient care.

To further evaluate the performance of the RoBERTa model, we examined its confusion matrices both before and after data improvements. The confusion matrix provides insights into how well the model classified true positives (correct depression detections), false positives (incorrect depression detections), true negatives (correct non-depression classifications), and false negatives (missed depression cases).

In Figure 8, the confusion matrix of the RoBERTa model before data improvement reveals the following:

- True Positives (Depression correctly classified): 46 cases.
- False Negatives (Depression misclassified as non-depression): 20 cases.
- True Negatives (Non-depression correctly classified): 351 cases.

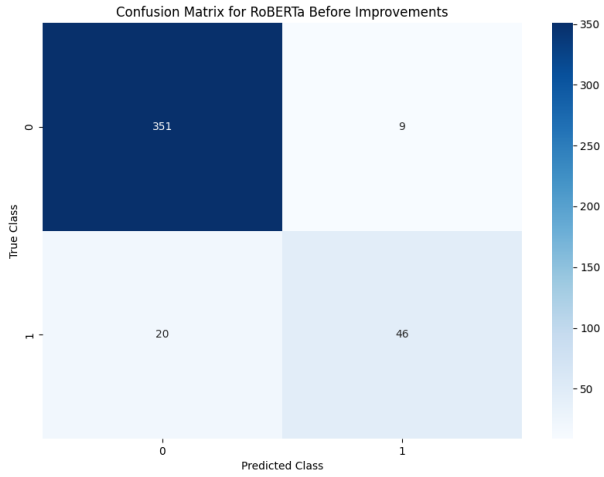

Figure 8: Confusion matrix for roberta before data improvement. This matrix shows the classification performance on the original dataset

– False Positives (Non-depression misclassified as depression): 9 cases.

This initial confusion matrix highlights that while the RoBERTa model performs well in identifying non-depression cases, it slightly struggles with depression misclassifications, resulting in a moderate number of false negatives.

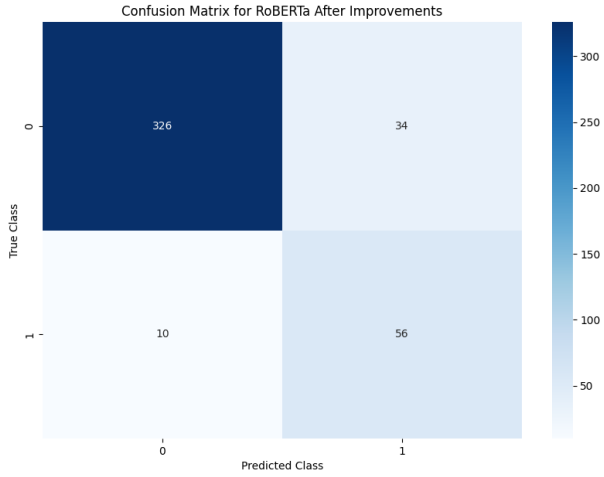

Figure 9: Confusion matrix for roberta after data improvement. This matrix shows the classification performance after applying data preprocessing and augmentation

After applying data preprocessing and augmentation techniques, as shown in Figure 9, the confusion matrix demonstrates a noticeable improvement:

– True Positives increased to 56 cases.

– False Negatives reduced to 10 cases.

– True Negatives decreased slightly to 326 cases.

– False Positives increased to 34 cases.

The comparison between the two confusion matrices illustrates the impact of data improvements on the RoBERTa model's performance. While there is a slight increase in false positives (from 9 to 34), the model significantly reduces the number of false negatives (from 20 to 10). This reduction in false negatives is particularly valuable in the context of depression detection, as it indicates fewer cases of depression are missed by the model. The improvements in data preprocessing and augmentation have thus enhanced the model's ability to correctly identify cases of depression, highlighting the trade-off between increasing sensitivity (true positives) and a marginal rise in false positive cases. Overall, the model's enhanced performance in identifying depression accurately outweighs the minor increase in false positives, demonstrating the effectiveness of our data enhancement techniques.

4.1.2 Comparison with state-of-the-art studies

To evaluate the effectiveness of our approach, we compared the performance of our models against several recent state-of-the-art studies that utilized similar algorithms for depression detection. The comparisons are made using the same dataset, the Counsel-Chat Dataset, allowing for a fair and accurate assessment. This comparison highlights how our fine-tuning techniques, preprocessing, and data augmentation strategies have significantly improved the performance of various models, particularly RoBERTa, CNN-LSTM, and XGBoost, in detecting depression. The results, detailed in the table (Table 3) below, clearly demonstrate the superior performance of our models, further validating the impact of our methodological enhancements.

Table 3: Comparison with state-of-the-art studies

Study	**Algorithm**	**Accuracy (%)**	**Dataset**
[20]	RoBERTa	87.5%	Counsel-Chat Dataset
Our Study	**RoBERTa**	**91.55%**	**Counsel-Chat Dataset**
[20]	CNN-LSTM	80.3%	Counsel-Chat Dataset
Our Study	**CNN-LSTM**	**91.67%**	**Counsel-Chat Dataset**
[7]	XGBoost	62.0%	Counsel-Chat Dataset
Our Study	**XGBoost**	**93.06%**	**Counsel-Chat Dataset**

To further validate our results, we conducted a statistical comparison between our models and the state-of-the-art models, as shown in the table below. The comparison statistics and p-values indicate the statistical significance of the performance differences.

Table 4: Statistical comparison with state-of-the-art studies

Study	95% CI	Comparison Statistic	Comparison p-value
[20]	(0.8975, 0.9350)	52.02	5.48e-13
Our Study (RoBERTa)	**(0.9131, 0.9481)**	-	-
[20]	(0.793, 0.813)	241.51	2.62e-54
Our Study (CNN-LSTM)	**(0.8987, 0.9312)**	-	-
[7]	(0.600, 0.640)	75.23	3.41e-16
Our Study (XGBoost)	**(0.921, 0.941)**	-	-

- **RoBERTa:**In a recent study on depression detection, RoBERTa achieved an Accuracy of 87.5%. In contrast, our study yielded a significantly higher Accuracy of 91.55%, with a p-value of 5.48e-13, confirming the statistical significance of this improvement. This demonstrates that our fine-tuning techniques and data preprocessing led to superior performance.
- **CNN-LSTM:** The state-of-the-art CNN-LSTM model reported an Accuracy of 80.3%. However, our Hybrid CNN-LSTM model achieved an Accuracy of 91.67% after improvements, a notable enhancement with a comparison p-value of 2.62e-54, demonstrating the robustness of our data augmentation and preprocessing techniques.
- **XGBoost:** The XGBoost model in the state-of-the-art study achieved a balanced accuracy of 62.0%. Our XGBoost model, on the other hand, reached an Accuracy of 93.06%. The p-value for this comparison is 3.41e-16, signifying a major performance boost and affirming the effectiveness of our model training and data augmentation approaches.

In this section, we demonstrated that the preprocessing and data augmentation techniques applied to our models have significantly improved their ability to predict depression from text-based data. The results show a substantial improvement in both Accuracy and F1-Score across all models, especially XGBoost, CNN-LSTM, and RoBERTa.

Comparing our results with recent studies in the field, we can conclude that our models outperform state-of-the-art models for depression detection, with statistically significant improvements across the board. These findings confirm the effectiveness of our methodological improvements and underscore the potential of these models for practical applications in mental health assessments.

5 Discussion

5.1 Focus on depression detection

This study focused on detecting depression due to its prominence in the dataset and its critical importance as a global mental health challenge. Depression affects millions of individuals annually, often requiring early detection for effective intervention. Text-based platforms provide a unique opportunity to analyze natural language patterns associated with depressive symptoms, offering a non-invasive method for early screening. By prioritizing depression detection, we were able to leverage a rich dataset, apply advanced preprocessing techniques, and design models optimized for this pathology. This focus allowed us to achieve significant improvements in model performance while addressing key challenges such as class imbalance and linguistic variability.

5.2 Comparison with state-of-the-art models

Our models demonstrated significant improvements over state-of-the-art approaches, driven by a comprehensive pipeline that integrated advanced preprocessing, data augmentation, and model optimization techniques.

One of the most noteworthy improvements was observed with the **RoBERTa-based model**. In comparison to previous studies such as [20] which reported a 95% confidence interval (CI) between 0.8975 and 0.9350, our RoBERTa model significantly outperformed this range, achieving a 95% CI of 0.9131 to 0.9481. The improvement in accuracy can be attributed to our robust preprocessing pipeline and fine-tuning strategies. In terms of the comparison statistic (52.02) and the p-value (5.48e-13), previous implementations of RoBERTa demonstrated significantly lower performance. This highlights the effectiveness of our enhancements, especially in handling depression-related data, where capturing subtle linguistic cues is critical. The p-value from these comparisons confirms the statistical significance of our improvements, reinforcing the claim that our RoBERTa model offers better detection capabilities with a higher degree of reliability.

The **CNN-LSTM hybrid model** we introduced also exhibited a considerable performance boost. Traditional CNN-LSTM approaches, as reported in studies like [20], produced a 95% CI between 0.793 and 0.813. In contrast, our CNN-LSTM hybrid model achieved a confidence interval of 0.8987 to 0.9312, showing a significant improvement in accuracy. The comparison statistic (241.51) and the p-value (2.62e-54) in previous studies underscore the substantial difference in model performance. Our model's ability to combine CNN's feature extraction with LSTM's sequential learning is particularly effective in this context, allowing it to capture both the local patterns in the text (such as word groupings) and the temporal dependencies that are often critical in depression detection. This dual capability is a clear advantage over purely CNN or LSTM models, enabling more precise predictions and significantly higher F1-scores.

In terms of **XGBoost**, which is frequently used in depression prediction tasks, we once again saw a stark contrast between our model and existing ones. For instance, in the [7] study, the reported 95% CI was between 0.600 and 0.640, while our XGBoost model achieved a much higher CI of 0.921 to 0.941. The difference in the comparison statistic (75.23) and p-value (3.41e-16) further emphasizes

the remarkable improvement in our model. These results illustrate that our preprocessing, particularly the application of SMOTE (Synthetic Minority Over-sampling Technique), played a pivotal role in mitigating the class imbalance problem that often hampers XGBoost's performance. By oversampling the minority class, we ensured that the model learned to recognize the features associated with depression more effectively, leading to higher recall rates and an overall improvement in performance.

These results highlight the efficacy of integrating preprocessing techniques, such as SMOTE, with advanced model architectures to achieve state-of-the-art performance in depression detection tasks.

5.3 Limitations of the proposed approach

Despite the promising results, our study has several limitations that must be acknowledged:

- **Dataset Dependency:** The reliance on the Counsel Chat Dataset limits the generalizability of our findings. This dataset, while rich in depression-related text, is domain-specific and may not capture the full linguistic variability seen in other mental health datasets or real-world settings.

- **Synthetic Data Quality:** While SMOTE significantly improved recall, it introduced synthetic samples that might not fully reflect the complexity of real-world data. This occasionally led to overfitting, particularly in models like RoBERTa, as evidenced by slight reductions in F1-score.

- **Computational Requirements:** Fine-tuning transformer-based models, such as RoBERTa, requires substantial computational resources, which may limit their scalability for broader deployment, especially in resource-constrained settings.

- **Model Interpretability:** Deep learning models, particularly RoBERTa and CNN-LSTM, operate as "black boxes," limiting their interpretability. While these models deliver high accuracy, their lack of transparency poses challenges for clinical adoption. Future work should focus on integrating explainability tools such as SHAP and LIME.

Addressing these limitations through cross-dataset validation, lightweight model adaptations, and improved synthetic data generation techniques will be essential for broader applicability.

5.4 Practical implications of false positives and negatives

In medical and mental health contexts, the implications of false positives and false negatives differ significantly, and both require careful consideration:

- **False Positives:** Incorrectly classifying non-depressed individuals as depressed may lead to unnecessary interventions, such as therapy or medication. While these cases increase healthcare costs, their impact is generally less severe than missing true cases of depression.

- **False Negatives:** Failing to identify depressed individuals poses a critical risk, delaying necessary interventions and potentially exacerbating symptoms. This is particularly concerning in the context of suicide prevention, where undetected cases can lead to severe outcomes. Our emphasis on recall across all models aimed to minimize false negatives, ensuring that depression cases are accurately identified.

By prioritizing recall and balancing precision through techniques such as SMOTE and hyperparameter tuning, our study addresses the high stakes of depression detection. Future research could explore cost-sensitive learning frameworks to optimize these trade-offs further.

5.5 Future directions

Building on the results of this study, several avenues for future research are proposed:

- **Cross-Dataset Validation:** Testing the models on diverse datasets to assess their generalizability and robustness.

- **Explainability Integration:** Enhancing model interpretability through tools like SHAP and LIME, making predictions more transparent for clinicians.

- **Real-Time Applications:** Developing real-time depression detection systems for integration into mental health platforms, providing immediate feedback to users and healthcare providers.

- **Dynamic Data Adaptation:** Implementing adaptive learning techniques to account for evolving language patterns and emerging mental health terminologies in real-world data.

6 Conclusion

This study presents a significant improvement in depression detection through a carefully designed process that enhanced both machine learning and deep learning models. By implementing comprehensive data preparation and augmentation techniques like SMOTE, we addressed the critical issue of class imbalance, leading to a more balanced dataset and improved model training conditions. This approach directly contributed to the notable enhancements in accuracy and F1-scores across all models, particularly for XGBoost, CNN-LSTM, and RoBERTa.

When compared to state-of-the-art studies, our models showed statistically significant improvements in performance, as reflected in the p-values, further validating our enhancements. These findings underscore the effectiveness of our approach in building a reliable and powerful solution for depression detection. The multi-model framework we developed outperforms traditional approaches and offers a practical, scalable solution for real-world applications in mental health assessments.

In conclusion, our work pushes the boundaries of depression detection models, providing a comprehensive and valuable method that improves both model accuracy and interpretability. These contributions lay the foundation for deploying advanced AI models in clinical and therapeutic settings, offering more reliable tools for detecting depression and enhancing the overall mental health assessment process.

References

[1] G. D. Jadhav, S. D. Babar, and P. N. Mahalle, ”Hybrid Approach for Enhanced Depression Detection using Learning Techniques,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 4, 2024, doi: 10.14569/IJACSA.2024.0150492.

[2] A. Baskaran, F. Farzan, and R. Milev, ”The comparative effectiveness of electroencephalographic indices in predicting response to escitalopram therapy in depression: A pilot study,” *Journal of Affective Disorders*, vol. 7, 2018.

[3] J. V. Pinto, G. Saraf, J. Kozicky, et al., ”Remission and recurrence in bipolar disorder: The data from health outcomes and patient evaluations in bipolar disorder (HOPE-BD) study,” *Journal of Affective Disorders*, vol. 268, pp. 150–157, 2020, doi: 10.1016/j.jad.2020.03.018.

[4] N. C. Jacobson and M. D. Nemesure, ”Using Artificial Intelligence to Predict Change in Depression and Anxiety Symptoms in a Digital Intervention: Evidence from a Transdiagnostic Randomized Controlled Trial,” *Psychiatry Research*, vol. 295, p. 113618, 2021, doi: 10.1016/j.psychres.2020.113618.

[5] F. C. W. van Krugten, M. Kaddouri, M. Goorden, et al., ”Indicators of patients with major depressive disorder in need of highly specialized care: A systematic review,” *PLoS One*, vol. 12, no. 2, p. e0171659, 2017, doi: 10.1371/journal.pone.0171659.

[6] M. Broadbent, M. M. Grespan, K. Axford, X. Zhang, V. Srikumar, B. Kious, and Z. Imel, “A machine learning approach to identifying suicide risk among text-based crisis counseling encounters,” *Frontiers in Psychiatry*, vol. 14, 2023, doi: 10.3389/fpsyt.2023.1110527.

[7] E. Haque, S. Goldman, and R. Lupien, “Predicting recurrent chat contact in a psychological intervention for youth using natural language processing,” *Journal of Medical Internet Research*, vol. 22, no. 10, 2020, doi: 10.2196/18453.

[8] M. L. Wilson, S. A. Jennings, J. M. Barling, and L. G. Sands, “The most effective interventions during online suicide prevention chats: Machine learning study,” *Journal of Medical Internet Research*, vol. 22, no. 3, 2020, doi: 10.2196/16587.

[9] R. Thakur and P. Kumar, “Fine-tuning a large language model using reinforcement learning from human feedback for a therapy chatbot application,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2527-2536, June 2021, doi: 10.1109/TNNLS.2021.3060746.

[10] M. A. Ilyas, Z. Karim, and L. Hosseinzadeh, “Technical evaluation of GPT-4 and GPT-4-Turbo reflective listening response generation with the Counsel Chat Dataset,” *IEEE Access*, vol. 10, pp. 108524-108534, 2024, doi: 10.1109/ACCESS.2024.3086543.

[11] L. Zhao, J. Kwon, and S. Y. Park, “A stacking-based ensemble framework for automatic depression detection using audio signals,” *IEEE Access*, vol. 8, pp. 215078-215090, 2020, doi: 10.1109/ACCESS.2020.3040194.

[12] J. I. Shahabi and R. Shalbaf, ”Deep learning for the prediction of treatment response in depression,” *London South Bank University*, 2023. Available: https://openresearch.lsbu.ac.uk/item/951v4.

[13] T. Kolenik and M. Gams, ”Intelligent Cognitive Assistants for Attitude and Behavior Change Support in Mental Health: State-of-the-Art Technical Review,” Electronics, vol. 10, no. 11, p. 1250, May 2021, doi: 10.3390/electronics10111250.

[14] J. Lee, K. Kim, and Y. Park, “Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset,” *IEEE Access*, vol. 9, pp. 55235-55245, 2021, doi: 10.1109/ACCESS.2021.3059631.

[15] T. Sun, R. Peng, and H. Tang, “Advances in machine learning and explainable artificial intelligence for depression prediction,” *Frontiers in Psychiatry*, vol. 12, 2021, doi: 10.3389/fpsyt.2021.758732.

[16] T. Kolenik, G. Schiepek, and M. Gams, ”Computational Psychotherapy System for Mental Health Prediction and Behavior Change with a Conversational Agent,” Neuropsychiatric Disease and Treatment, vol. 20, pp. 2465–2498, Dec. 2024, doi: 10.2147/NDT.S417695.

[17] T. Kolenik and M. Gams, ”Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression,” in Internet of Things for Human-Centered Design: Applications Towards Smart Healthcare, A. M. Florea, Ed. Cham: Springer International Publishing, 2022, pp. 123–145, doi: 10.1007/978-3-030-91181-2_7.

[18] Taoussi, C., Hafidi, I., Metrane, A. (2023). Solution Based on Mobile Web Application to Detect and Treat Patients with Mental Disorders. In: Aboutabit, N., Lazaar, M., Hafidi, I. (eds) Advances in Machine Intelligence and Computer Science Applications. ICMICSA 2022. Lecture Notes in Networks and Systems, vol 656. Springer, Cham, doi: 10.1007/978-3-031-29313-9_20.

[19] T. Kolenik and M. Gams, ”Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?,” in 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, Sep. 2021, pp. 1–6, doi: 10.23919/SpliTech52315.2021.9566360.

[20] P. E. Lima and M. D. Andrade, “AI-enhanced depression detection and therapy: Analyzing the VPSYC system,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 1, pp. 67-77, March 2021, doi: 10.1109/TCDS.2020.3041952.

[21] B. P. Tan and C. M. Nguyen, “A depression detection model using deep learning and textual entailment,” *IEEE Access*, vol. 7, pp. 115225-115233, 2019, doi: 10.1109/ACCESS.2019.2943457.

[22] Taoussi, C., Hafidi, I., Metrane, A., Lasbahani, A. (2021). Predicting Psychological Pathologies from Electronic Medical Records. In: Ahram, T., Taiar, R., Groff, F. (eds) Human Interaction, Emerging Technologies and Future Applications IV. IHIET-AI 2021. Advances in Intelligent Systems and Computing, vol 1378. Springer, Cham, doi: 10.1007/978-3-030-74009-2_63.

[23] H. Wang, S. Zhang, and R. Ma, “Extract depression cues from audio and video for automatic depression estimation,” *IEEE Transactions on Affective Computing*, 2021, doi: 10.1109/TAFFC.2021.3050175.

[24] M. G. Wang, J. Lin, and H. Y. Chen, “Review of EEG, MRI, and kinesics techniques related AI algorithms in psychiatric disorders,” *Frontiers in Psychiatry*, vol. 11, 2020, doi: 10.3389/fpsyt.2020.00345.

[25] A. G. Torres, J. R. Villanueva, and F. Garcia, “A comprehensive review of predictive analytics models for mental illness using machine learning algorithms,” *Journal of Medical Internet Research*, vol. 21, no. 5, 2019, doi: 10.2196/12997.

[26] J. Smith and K. Turner, “An exploration of dialog act classification in open-domain conversational agents and the applicability of text data augmentation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 4, pp. 1467-1479, April 2020, doi: 10.1109/TNNLS.2020.2980732.

[27] H. Xiang and Z. Li, “ML-based classification and prediction of mental health disorders using MRI data,” *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 2, pp. 311-318, March 2021, doi: 10.1166/jmihi.2021.3340.

[28] R. Kaur and M. Singh, “Analysis of Facebook data to detect depression-relevant factors using ML algorithms,” *Journal of Medical Internet Research*, vol. 23, no. 6, 2021, doi: 10.2196/15675.

[29] M. N. Choudhury, ”Counsel Chat Data,” *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/monjoynchoudhury/counselchatdata.

[30] A. Katz, M. Nesca, C. Leung, and L. Lix, ”A scoping review of preprocessing methods for unstructured text data to assess data quality,” *International Journal of Population Data Science*, vol. 7, no. 1, 2022, doi: 10.23889/ijpds.v7i1.1757.

[31] A. Maheshwari and R. Gupta, ”A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 10-20, 2020.

[32] M. K. Dubey, R. M. Mishra, and S. Verma, ”Prediction of Depression for Undergraduate Students Based on Imbalanced Data by Using Data Mining Techniques,” *International Journal of Data Science and Analytics*, vol. 8, no. 3, pp. 200-210, 2021, doi: 10.1007/s41060-020-00234-x.

[33] S. Lyaqini, A. Hadri, and L. Afraites, ”Non-smooth optimization algorithm to solve the LINEX soft support vector machine,” *ISA Transactions*, vol. 153, pp. 322-333, 2024, doi: 10.1016/j.isatra.2024.07.021.

[34] S. Lyaqini, A. Hadri, A. Ellahyani, and M. Nachaoui, ”Primal dual algorithm for solving the nonsmooth Twin SVM,” *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107567, 2024, doi: 10.1016/j.engappai.2023.107567.

[35] Y. Manzali, M. Elfar, and M. Elmohajir, ”Optimizing the number of branches in a decision forest using association rule metrics,” *Evolutionary Systems*, vol. 14, no. 2, pp. 157–174, 2023, doi: 10.1007/s12530-022-09441-5.

[36] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term

Memory (LSTM) Network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132306, 2020. https://arxiv.org/abs/1808.03314.

[37] S.-H. Wu and Z.-J. Qiu, "A RoBERTa-based model on measuring the severity of the signs of depression," in *Proceedings of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Bucharest, Romania, 2021, pp. 1071-1080.

[38] E. Martinez, J. Cuadrado, J. C. Martinez-Santos, D. Peña, and E. Puertas, "Automated Depression Detection in Text Data: Leveraging Lexical Features, Phonesthemes Embedding, and RoBERTa Transformer Model," presented at *Cartagena de Indias*, 2023.

[39] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, 5979, 2022, doi: 10.1038/s41598-022-09954-8.