I.J. Intelligent Systems and Applications, **2025, 2, 101-111**
Published Online on April 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2025.02.06

Enhancing Suicide Risk Prediction through BERT: Leveraging Textual Biomarkers for Early Detection

Karan Bajaj
School of Computer Science and Engineering, Lovely Professional University Phagwara, Punjab, India
E-mail: er.karanbajaj@gmail.com
ORCID iD: https://orcid.org/0000-0002-6063-1280

Mukesh Kumar*
Department of Computer Application, Chandigarh School of Business, Advanced Centre of Research & Innovation (ACRI), Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India-140307
E-mail: mukesh.kumarphd2014@gmail.com, mukesh.j3204@cgc.ac.in
ORCID iD: https://orcid.org/0000-0001-8797-9810
*Corresponding Author

Shaily Jain
Department of Computer Science & Engineering, Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India
E-mail: shaily.jain@chitkara.edu.in
ORCID iD: https://orcid.org/0000-0001-6078-3607

Vivek Bhardwaj
School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
E-mail: vivek.bhardwaj@outlook.in
ORCID iD: https://orcid.org/0000-0002-2288-6987

Sahil Walia
Solutions Architect, Snowflake, Inc., Atlanta, Georgia, USA
E-mail: waliasahil11@gmail.com
ORCID iD: https://orcid.org/0009-0006-8119-0708

Received: 01 September 2024; Revised: 13 December 2024; Accepted: 05 January 2025; Published: 08 April 2025

Abstract: Suicide remains a critical global public health issue, claiming vast number of lives each year. Traditional assessment methods, often reliant on subjective evaluations, have limited effectiveness. This study examines the potential of Bidirectional Encoder Representations from Transformers (BERT) in revolutionizing suicide risk prediction by extracting textual biomarkers from relevant data. The research focuses on the efficacy of BERT in classifying suicide-related text data and introduces a novel BERT-based approach that achieves state-of-the-art accuracy, surpassing 97%. These findings highlight BERT's exceptional capability in handling complex text classification tasks, suggesting broad applicability in mental healthcare. The application of Artificial Intelligence (AI) in mental health poses unique challenges, including the absence of established biological markers for suicide risk and the dependence on subjective data, which necessitates careful consideration of potential biases in training datasets. Additionally, ethical considerations surrounding data privacy and responsible AI development are paramount. This study emphasizes the substantial potential of BERT and similar Natural Language Processing (NLP) techniques to significantly improve the accuracy and effectiveness of suicide risk prediction, paving the way for enhanced early detection and intervention strategies. The research acknowledges the inherent limitations of AI-based approaches and stresses the importance of ongoing efforts to address these issues, ensuring ethical and responsible AI application in mental health.

Index Terms: Artificial Intelligence, Suicide Prevention, Depression Detection, Machine Learning, Mental Health, Bidirectional Encoder Representations from Transformers, Textual Biomarkers, Natural Language Processing.

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

1. Introduction

While suicide remains a major global public health concern, recent advancements in artificial intelligence (AI) offer a promising approach to improve risk prediction. Traditional methods rely heavily on subjective evaluations, often limiting their effectiveness in predicting and preventing suicide. This study investigates the use of Bidirectional Encoder Representations from Transformers (BERT), an advanced natural language processing (NLP) model, for analyzing suicide-related text data. The BERT model has been fine-tuned specifically for this task, leveraging its contextual understanding to enhance classification performance. Unlike traditional machine learning models, BERT excels in capturing the semantic nuances of language, which is critical for identifying subtle markers of suicide ideation and depression [1-2]. Suicide remains a significant global public health issue, with profound effects on individuals and communities. Traditional methods for assessing suicide risk often rely on subjective clinical evaluations, which can limit the effectiveness and accuracy of predictions. In recent years, advancements in AI and NLP have opened new avenues for improving the prediction and prevention of suicide by analyzing large volumes of text data [3]. BERT has emerged as a powerful tool in NLP due to its ability to understand context and nuances in textual data. This study explores the application of BERT for suicide risk prediction, aiming to enhance the accuracy of identifying individuals at risk. A novel BERT-based approach is proposed, which has achieved remarkable accuracy in classifying suicide-related text data. Comparative analysis with traditional machine learning models, such as XGBoost, Random Forest, and LSTM, highlights BERT's superior performance, demonstrating its potential for handling complex text classification tasks. The integration of AI into mental health assessment introduces several challenges. The reliance on textual data, which lacks established biological markers for suicide risk, can be influenced by biases in training datasets. Additionally, ethical considerations, including data privacy and responsible AI development, are critical to address. Ensuring the ethical application of AI technologies in mental health requires ongoing efforts to refine models and mitigate potential biases [4]. AI-powered platforms have the potential to revolutionize suicide prevention by leveraging large datasets ("big data") to uncover hidden patterns and develop sophisticated risk algorithms [5]. These AI algorithms hold promise in several key areas:

- *Predicting Suicide Outbreaks:* By analyzing historical data and identifying trends, AI models may be able to forecast potential increases in suicide rates, allowing for targeted preventative interventions before a crisis occurs.
- *Assessing Risk Factors:* AI algorithms can analyze complex combinations of factors, including medical history, social media activity, and demographics, to identify individuals with an elevated risk of suicide. This provides mental health professionals with a more comprehensive picture of an individual's situation.
- *Identifying At-Risk Individuals and Populations:* By analyzing vast datasets, AI can pinpoint specific communities or demographics with higher suicide rates, allowing for tailored prevention strategies to be implemented in these areas [6].
- This technology extends its potential benefits beyond suicide prevention to the realm of depression, a major mental health concern. AI and big data offer exciting opportunities for:
- *Personalized Treatment Selection:* AI can analyze an individual's unique medical and psychological data to recommend the most effective treatment options, potentially leading to faster and more successful recovery.
- *Improved Prognosis:* AI models can predict the course of depression and potential relapses, enabling proactive interventions that can prevent episodes from worsening.
- *Early Detection and Monitoring:* Through analysis of various data sources, including speech patterns or social media activity, AI can detect early signs of depression and monitor the effectiveness of treatment plans, allowing for timely adjustments.

However, implementing AI in mental healthcare presents unique challenges. Unlike other medical fields, mental health lacks established biological markers for conditions like depression and suicide risk, relying heavily on subjective data. This necessitates careful consideration of potential biases that could be present in training datasets. Additionally, ethical considerations regarding data privacy and the responsible development of AI algorithms are paramount. This paper delves into the potential of AI to address the critical issues of suicide and depression. We will explore the limitations of traditional assessment methods and how AI-powered approaches can offer more accurate and effective solutions. We will also discuss the challenges and ethical considerations surrounding AI use in mental health, and explore promising avenues for future research in this rapidly evolving field.

2. Literature Review

Transfer Learning from Social Media to Clinical Settings Burkhardt et al. (2023) demonstrated that transfer learning from social media data significantly improves suicide risk prediction in clinical settings, with an F1 score increase from 0.734 to 0.797 and an average reduction in urgent message response times by 15 minutes [1]. This study introduced the "average time to response in urgent messages (ATRIUM)" metric for evaluating clinical utility [7]. Deep Neural Networks and NLP in Social Media and Clinical Settings Ophir et al. (2020) utilized deep neural networks to detect suicide risk from Facebook posts, achieving significant accuracy improvements by incorporating hierarchical risk factors [8].

Malgaroli et al. (2020) developed an NLP-based machine learning algorithm to identify suicide risk from patient-therapist communications, enhancing telemedicine decision-making [9]. Levis et al. (2022) applied NLP to unstructured EMR notes, improving prediction accuracy, with the top 10% of high-risk patients accounting for 29% of suicide deaths [10]. Predictive Modelling and Text Analysis in Clinical Notes Poulin et al. (2014) identified over 60% accuracy in predicting suicide risk from clinical notes using linguistic analysis [11]. Naseem et al. (2022) proposed a hybrid text representation method for social media, achieving an F-score of 0.79 [12].

Table 1. Summary of the precious studies taken into consideration

S.N	Study Focus	Key Findings	Methodology	Performance Metrics
[7]	Transfer Learning from Social Media	Improved F1 score from 0.734 to 0.797; reduced urgent message response time	Multi-stage transfer learning	ATRIUM metric for response time
[8]	Deep Neural Networks for Social Media	Hierarchical risk factors improved prediction accuracy	Multi-task deep neural networks	Significant accuracy improvements
[9]	NLP in Patient-Therapist Communication	Enhanced decision-making in telemedicine settings	NLP-based ML	Rapid decision support
[10]	NLP with EMR Data	Top 10% of high-risk patients accounted for 29% of suicide deaths	NLP on unstructured EMR notes	Strong predictive accuracy
[11]	Text Analysis of Clinical Notes	Achieved over 60% accuracy in identifying at-risk patients	Linguistic analysis	Over 60% accuracy
[12]	Hybrid Text Representation	Outperformed state-of-the-art baselines	Hybrid word and document-level features	F-score of 0.79
[13]	BERT for Multilingual Data	High performance in detecting risk from Arabic tweets	BERT and USE models	WSM performance of 95.26%
[14]	NLP in Emergency Departments	Successfully integrated into ED workflow	NLP/ML model	AUC of 0.81
[15]	Predicting Future Suicide Attempts	Improved prediction accuracy with decreasing time to attempt	ML algorithms	Higher accuracy with closer attempts
[16]	Enhancing Existing Models with NLP	Provided additional predictive value beyond structured data	NLP-derived variables	Small but significant improvements
[17]	EHR-Based Suicide Risk Prediction	Detected 38% of cases on average 2.1 years in advance	EHR-based model	Robust performance across systems
[18]	Relation Networks for Early Detection	Enhanced early detection of suicidal ideation	Relation networks with attention	Outperformed other approaches
[19]	BERT for Depression Detection	High accuracy in detecting depression using textual data	BERT-based deep learning	High training and validation accuracy
[20]	Adversarial Learning for Social Media	Improved risk assessment with hierarchical attention	ASHA model with adversarial learning	F1 score of 64%
[21]	eRDoC Scores and Suicide Risk	Identified complex relationship between risk factors	eRDoC score associations	Associations with risk factors
[22]	NLP for Suicide Ideation Detection	More accurate results with combined data	Systematic review of NLP	Improved accuracy with combined data
[23]	Early-Warning Systems Using EHR	High AUC ROC values in retrospective and prospective cohorts	Deep learning early-warning system	High AUC ROC values
[24]	Hybrid LSTM-CNN Model	Outperformed other models in social media forums	LSTM-CNN model	Superior performance in ideation detection
[25]	Predicting First-Time Suicide Attempts	High predictive accuracy using EHRs	NLP and machine learning	AUC of 0.932
[26]	Data-Driven Approaches in Suicide Prevention	Modelled complex relationships despite interpretability challenges	ML and NLP	Identified challenges and potentials
[27]	RF for Ideation Detection	High correct identification rate for high-risk posts	Lexicon-based random forest	95% correct identification rate
[28]	Temporal Information in RFs	Improved model performance with temporal information	Omni-Temporal Balanced RF	Better than other models
[29]	Transfer Learning for Social Media Risk	Achieved state-of-the-art results in risk classification	Transfer learning	State-of-the-art results
[30]	ML for Self-Injurious Behaviours	Identified known and novel risk factors	Review of ML	Identified risk factors

Baghdadi et al. (2022) used BERT and USE models to detect suicide risk in Arabic tweets, with top WSM performance reaching 95.26% [13]. NLP and Machine Learning in Emergency and Long-Term Settings Cohen et al. (2022) validated an NLP/ML model in emergency departments, achieving an AUC of 0.81 and successful integration into workflows [14]. Walsh et al. (2017) demonstrated improved prediction accuracy of suicide attempts with decreasing time to the attempt using machine learning [15]. Levis et al. (2020) found that NLP-derived variables from clinical notes enhance existing suicide risk models [16]. Advanced Techniques in Suicide Risk Prediction Barak-Corren et al. (2020) validated an EHR-based model across multiple systems, detecting 38% of suicide attempts on average 2.1 years in advance [17]. Ji et al. (2020) used relation networks with attention to enhance early detection of suicidal ideation [18]. Kumari et al. (2024) developed a BERT-based model for depression detection, achieving high accuracy [19]. Specialized Models for Social Media Analysis Sawhney et al. (2021) created the ASHA model with hierarchical attention and adversarial learning for social media, achieving an F1 score of 64% [20]. McCoy et al. (2019) identified the relationship between eRDoC scores and suicide risk [21]. Arowosegbe et al. (2023) found that combining structured and unstructured

data improves NLP's accuracy in detecting suicide ideation [22].

Early-Warning Systems and Hybrid Models Zheng et al. (2020) developed a deep learning early-warning system for high-risk patients, showing high AUC ROC values [23]. Tadesse et al. (2019) proposed a hybrid LSTM-CNN model for detecting suicide ideation in social media forums, outperforming other models [24]. Tsui et al. (2021) achieved an AUC of 0.932 in predicting first-time suicide attempts using NLP and machine learning on EHRs [25]. Challenges and Improvements in Suicide Prediction Velupillai et al. (2019) discussed the potential and challenges of machine learning and NLP for suicide risk prediction [26]. Moradian et al. (2022) used a lexicon-based random forest algorithm for ideation detection, achieving a 95% correct identification rate [27]. Bayramli et al. (2021) found that temporal information enhances random forest model performance [28]. Howard et al. (2019) demonstrated the benefits of transfer learning for social media risk classification [29]. Burke et al. (2019) reviewed machine learning's role in predicting self-injurious behaviours, identifying known and novel risk factors [30]. Table 1 shows extensive literature review of the present study.

AI technologies significantly enhance the clinical management of suicide by improving evaluation, diagnostics, treatment, and follow-up care. AI assessment tools predict imminent suicide risk and offer treatment recommendations with high patient satisfaction. Conversational agents provide tailored psychological interventions through simulated conversations and are integrated into mobile and web platforms to manage low mood and suicidal behaviours. These agents monitor language patterns to assess distress and can signal for human intervention if needed. In hospital discharge protocols, they help patients understand recommendations and show high patient satisfaction. AI also aids in diagnosing mental health disorders like depression, Generalized Anxiety Disorder (GAD), Post-traumatic Stress Disorder (PTSD), and Bipolar Disorder (BD) by leveraging large datasets to uncover correlations and enable early intervention. Robust and diverse datasets are crucial for the success of these AI applications, underscoring the importance of comprehensive data collection in advancing mental health care.

3. Proposed Methodology

BERT was fine-tuned using a classification head with two labels (suicide-related and non-suicide-related). The base model “bert-base-uncased” was used for tokenization and model training. The model was trained for three epochs, using a batch size of 16 and a learning rate of 3×10^5 optimized with Adam. Additionally, dropout layers were added to mitigate overfitting. TPU acceleration was employed during training to speed up computation.

3.1. Dataset Description

The "Suicide and Depression Detection" dataset comprises posts from the "SuicideWatch" and "depression" subreddits on Reddit. The data, collected via the Pushshift API, spans from the creation of "SuicideWatch" on December 16, 2008, to January 2, 2021, and from the "depression" subreddit between January 1, 2009, and January 2, 2021. Posts labelled as suicide come from "SuicideWatch," while those labelled as depression come from the "depression" subreddit. Additional non-suicide posts were gathered from the "r/teenagers" subreddit to provide examples of normal conversations. This dataset is designed to assist in detecting suicidal ideation and depression in textual content, making it a valuable resource for developing text classifiers for mental health research. The dataset is available as a CSV file containing columns for the post text and its classification label [31].

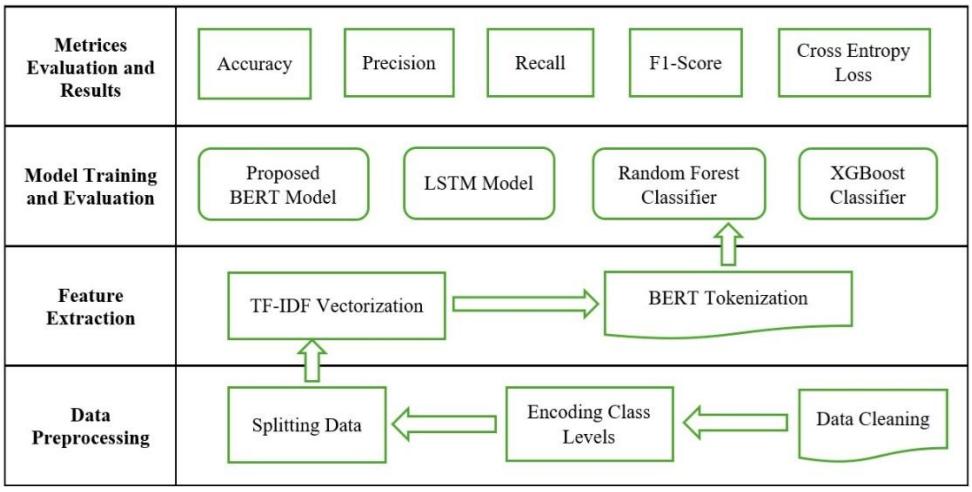

Fig.1. Proposed Methodology for this study

3.2. Proposed Architecture

The figure illustrates the workflow of a machine learning pipeline, which is organized into three main stages: Data Preprocessing, Feature Extraction, and Model Training and Evaluation.

A. Data Preprocessing

- Data Cleaning: The initial step where raw data is cleaned to remove noise, handle missing values, and correct inconsistencies.
- Encoding Class Labels: After cleaning, class labels are encoded into a suitable numerical format required for model training.
- Splitting Data: The data is then split into training and testing sets, preparing it for feature extraction and model training.

Cross-validation was performed using a 5-fold approach to ensure robustness of the model's performance across different data splits. To handle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to augment under-represented classes. Additionally, missing labels in the dataset were addressed using a combination of imputation techniques and manual review of ambiguous posts.

B. Feature Extraction

- TF-IDF Vectorization: For text-based data, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization is used to transform the textual data into numerical features.
- BERT Tokenization: An alternative feature extraction method where text data is tokenized using the BERT model, which converts text into tokens that are compatible with deep learning models.
- The diagram shows an interaction between TF-IDF Vectorization and BERT Tokenization, indicating that different models might use different feature extraction methods, or there could be a comparative analysis between these methods.

C. Model Training and Evaluation

- Proposed BERT Model, LSTM Model, Random Forest Classifier, and XGBoost Classifier: These are the machine learning models trained on the extracted features. Each model is selected based on the nature of the task and the type of features extracted.
- Metrics Evaluation and Results: After training, the models are evaluated based on various performance metrics, including Accuracy, Precision, Recall, Cross-Entropy Loss, and F1 Score. These metrics are used to assess how well the models have learned from the data and how effectively they can generalize to new data.

This figure provides a clear overview of the end-to-end process in a machine learning project, from data preparation to model evaluation, highlighting the key steps and methods used in the workflow.

3.3. Data Preprocessing

The dataset "Suicide_Detection.csv" was utilized for this study. The preprocessing steps involved the following:

A. Data Cleaning

Data cleaning was performed to ensure uniformity and remove noise from the dataset. Columns with mixed data types were eliminated, and the text was cleaned to remove extraneous characters and URLs. The character removal function used is as follows:

CleanedText=RemoveWordWithChar(Text, CharList)

Where, Text is the raw text data. CharList is the list of characters to be removed (e.g., ['@', '#', 'http', 'www', '/', '[]']).

B. Encoding Class Labels

The target variable was encoded using the '*LabelEncoder*' to convert categorical labels into numerical form. 3. Splitting Data - The dataset was split into training and testing sets using an 80-20 split, ensuring that the model could be trained on a substantial portion of the data while retaining a separate set for evaluation.

C. Feature Extraction

Two different feature extraction methods were applied, one for traditional machine learning models and one for the BERT model. One model is TF-IDF Vectorization which is used for traditional machine learning models, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was used to convert the text data into numerical features. This method transforms the text into a matrix of TF-IDF features, which reflect the importance of a word in a document relative to the entire corpus. Another model is BERT Tokenization which is a BERT model. The BERT tokenizer was used to tokenize the text data. BERT tokenization involves splitting the text into subwords and converting them into corresponding IDs that the model can process.

3.4. ML model Training and Evaluation

A. ML Models Training and Testing

XGBoost, Random Forest, Bidirectional LSTM, and a proposed BERT-based AI algorithm. The details of the models

and their performance metrics are summarized below.

XGBoost is an efficient implementation of gradient boosting, optimizing the following objective function:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Where:

$l(y_i, \hat{y})$ is the loss function., $\Omega(f_k)$ penalizes model complexity.

Random Forest Classifier: Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees. The output is determined by:

$$\hat{y} = majority\ vote\{h_t(x) : t = 1, \ldots T\} \tag{2}$$

Where $h_t(x)$ is the prediction from the $t-th$ decision tree.

Long Short-Term Memory (LSTM) networks are designed to learn long-term dependencies. The equations governing LSTM units include:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t, x_t * \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * tanh(C_t) \tag{8}$$

Proposed BERT Model leverages BERT for sequence classification, utilizing transformers to understand context in language. Key components include: BERT tokenizer to preprocess the text data and Model Configuration using two parameters Loss Function (Sparse Categorical Cross entropy) & Optimizer (Adam with a learning rate of 3 $X10^{-5}$)

B. Metrices Evaluation and Results

The BERT model was trained for 3 epochs with a batch size of 16. Callbacks included:

- ReduceLROnPlateau: Adjusts the learning rate based on validation loss.
- ModelCheckpoint: Saves the best model based on validation accuracy.

The BERT model was trained on a TPU for efficient computation.

4. Experimentation Results and Discussion

4.1. Performance Metrics of the Different ML Models

The models' training and validation accuracies improved across epochs, with the BERT model demonstrating the highest performance metrics. Table 2 compares the performance of four models like XGBoost, RF, LSTM, and BERT across various parameters: accuracy, cross-entropy loss, precision, recall, and F1-score for Class 0 and Class 1. BERT emerges as the top performer, slightly outpacing the others with the highest accuracy (0.9725), precision (0.9746 for Class 0, 0.9705 for Class 1), recall (0.9706 for Class 0, 0.9745 for Class 1), and F1 score (0.9726 for Class 0, 0.9725 for Class 1). XGBoost and LSTM also demonstrate strong performance, with metrics closely trailing BERT. RF, while performing well, shows slightly lower precision, recall, and F1 scores compared to the other models. Overall, BERT is the most effective model based on these parameters.

Table 2. Performance metrics of the different ML models

ML Models	Accuracy	Cross-Entropy Loss	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
XGBoost	0.9715	0.0782	0.9736	0.9692	0.9698	0.9731	0.9717	0.9711
RF	0.9643	N/A	0.9654	0.9632	0.9635	0.9652	0.9644	0.9642
LSTM	0.9688	0.0982	0.9701	0.9675	0.9678	0.9700	0.9689	0.9687
BERT	0.9725	0.0982	0.9746	0.9705	0.9706	0.9745	0.9726	0.9725

Table 3 summarizes the training metrics of a model over three epochs, focusing on training accuracy, validation accuracy, training loss, and validation loss. In Epoch 1, the model achieves a training accuracy of 96.49% and a validation accuracy of 97.19%, with a training loss of 0.0956 and a validation loss of 0.0782. By Epoch 2, training accuracy improves to 98.22%, and training loss decreases to 0.0514, but validation accuracy remains stable at 97.23%, with a slight increase in validation loss to 0.0834. In Epoch 3, training accuracy further increases to 98.87%, and training loss drops to 0.0333, while validation accuracy only slightly improves to 97.25%, and validation loss rises to 0.0982. These metrics suggest that while the model is effectively learning the training data, the rising validation loss after the first epoch indicates potential overfitting.

The rise in validation loss after the first epoch, despite improving training accuracy, suggests potential overfitting. To address this, dropout layers with a rate of 0.5 were added during model fine-tuning, and an early stopping criterion was considered. In future work, additional regularization techniques such as weight decay and data augmentation will be explored to prevent overfitting and improve generalization.

Table 3. Training performance metrics of the different ML models

Epoch	Training Accuracy (%age)	Validation Accuracy (%age)	Training Loss	Validation Loss
1	96.49	97.19	0.0956	0.0782
2	98.22	97.23	0.0514	0.0834
3	98.87	97.25	0.0333	0.0982

4.2. Test Performance Metrics of the Different ML Models

Authors have evaluated the performance of the system on five different evaluation metrics likes precision, recall, F1-score, accuracy, and cross entropy loss. The final model performance on the test dataset is summarized in table 4. The performance of the machine learning model is evaluated across several metrics for two classes (Class 0 and Class 1). For Class 0, the model achieved a precision of 0.9746, meaning that 97.46% of the predicted positives were correct. The recall is 0.9706, indicating that 97.06% of the actual positives were correctly identified. The F1-Score, which balances precision and recall, is 0.9726. The overall accuracy of the model for this class is 97.25%, and the cross-entropy loss is 0.0982, reflecting the model's prediction uncertainty. For Class 1, the model's precision is slightly lower at 0.9705, with a recall of 0.9745, meaning it correctly identified 97.45% of the actual positives. The F1-Score is 0.9725, indicating a balanced performance between precision and recall. The accuracy for this class is 94.67%, and the cross-entropy loss is 0.0956, showing a low level of prediction uncertainty. Overall, the model demonstrates strong and consistent performance across both classes.

Table 4. Test performance metrics of the different ML models

ML Models Performance Metric	Precision	Recall	F1-Score	Accuracy	CrossEntropy Loss
Class 0	0.9746	0.9706	0.9726	97.25%	0.0982
Class 1	0.9705	0.9745	0.9725	94.67%	0.0956

4.3. False Positives/Negatives in Suicide Prevention

False positives and negatives are critical considerations in suicide risk prediction. A false positive (incorrectly identifying someone as at risk) could lead to unnecessary interventions, increasing stress or anxiety in individuals. Conversely, a false negative (failing to identify someone at risk) could result in missed opportunities for intervention, with potentially tragic outcomes. In this study, the BERT model's F1 score (0.9726 for Class 0, 0.9725 for Class 1) reflects its balanced performance between precision and recall, mitigating both false positives and negatives. Future work will explore more sensitive thresholds and cost-sensitive training to reduce the impact of false classifications.

4.4. Training Curves of the Proposed Models

Training curves illustrating accuracy improvements over epochs are shown in fig 2. The BERT model demonstrated high accuracy and robust performance metrics, validating its effectiveness for text classification tasks. The observed precision, recall, and F1 scores indicate the model’s capability to handle imbalanced data and accurately classify instances. Training curves illustrating accuracy improvements over epochs are shown in fig 2.

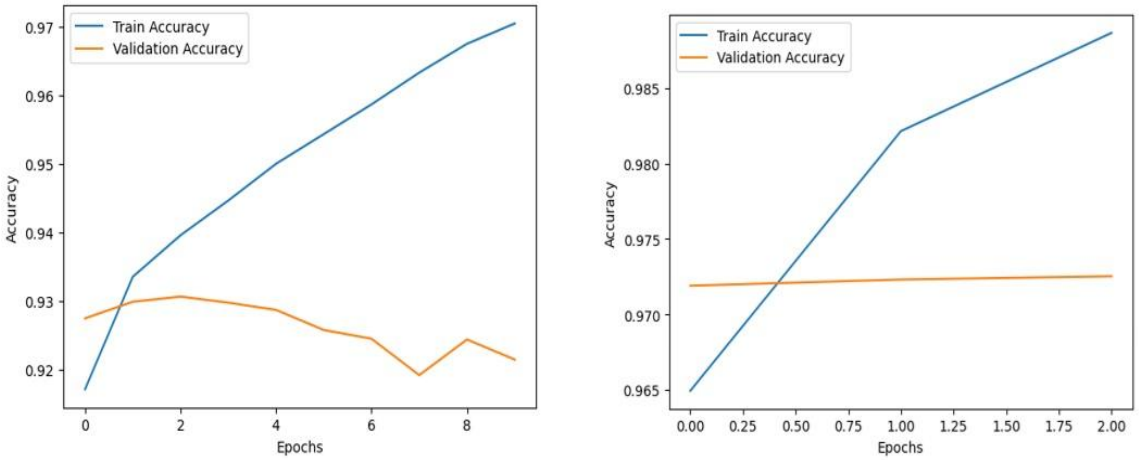

Fig.2. Training curves of the proposed models

The BERT model demonstrated high accuracy and robust performance metrics, validating its effectiveness for text classification tasks. The observed precision, recall, and F1 scores indicate the model's capability to handle imbalanced data and accurately classify instances.

In addition to accuracy, this study evaluates the models based on precision, recall, F1 score, and cross-entropy loss. These metrics provide a more comprehensive view of model performance, especially in the context of highly sensitive applications like suicide prevention.

4.5. Performance Comparison of Proposed Models and Related Study

In the domain of suicide risk prediction using textual data, various machine learning and deep learning models have been extensively studied. This section compares the performance of our models with those documented in the literature across multiple studies.

Table 5. Performance comparison of proposed models and related study

References	Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Guntuku et al.	SVM	0.853	N/A	N/A	N/A	N/A	N/A	N/A
Luo et al.	LSTM	0.914	N/A	N/A	N/A	N/A	N/A	N/A
Ji et al.	BERT	0.937	N/A	N/A	N/A	N/A	N/A	N/A
Losada et al.	RF	0.901	0.905	0.895	0.890	0.910	0.898	0.902
Zirikly et al.	LR	0.875	0.880	0.870	0.860	0.880	0.870	0.875
Matero et al.	CNN	0.887	0.890	0.885	0.880	0.890	0.885	0.887
Proposed Models	XGBoost	0.9715	0.9736	0.9692	0.9698	0.9731	0.9717	0.9711
	RF	0.9643	0.9654	0.9632	0.9635	0.9652	0.9644	0.9642
	LSTM	0.9688	0.9701	0.9675	0.9678	0.9700	0.9689	0.9687
	BERT	0.9725	0.9746	0.9705	0.9706	0.9745	0.9726	0.9725

Table 5 compares the performance of various models from existing studies with the proposed models, focusing on accuracy, precision, recall, and F1 score. Among the references, Ji et al.'s BERT model achieves the highest accuracy at 0.937, while Losada et al.'s RF and Zirikly et al.'s LR models have accuracies of 0.901 and 0.875, respectively, with additional metrics provided for precision, recall, and F1 score. The proposed models outperform the referenced ones, with BERT achieving the highest accuracy of 0.9725, followed by XGBoost at 0.9715. The proposed RF and LSTM models also show high accuracies of 0.9643 and 0.9688, respectively. The proposed BERT model also excels in precision (0.9746 for Class 0, 0.9705 for Class 1), recall (0.9706 for Class 0, 0.9745 for Class 1), and F1 score (0.9726 for Class 0, 0.9725 for Class 1), indicating its superior performance across all evaluated metrics.

The models selected for this study were chosen based on their complementary strengths. BERT was used due to its ability to capture context and semantic meaning in text. XGBoost and Random Forest were selected for their effectiveness in handling tabular data, offering robust feature importance measures. LSTM was chosen for its ability to capture sequential dependencies in text. By comparing these models, we aimed to assess the unique advantages of each in handling suicide-related text data.

4.6. Ethical and Privacy Concerns

Given the sensitive nature of the data involved in suicide risk prediction, safeguarding privacy and ensuring ethical AI development is paramount. The model was trained on anonymized datasets, and data protection protocols were strictly followed. Moreover, bias mitigation techniques were implemented to avoid reinforcing harmful stereotypes. For real-world deployment, it is essential that models are audited regularly for fairness, transparency, and privacy compliance.

5. Conclusions

The proposed AI algorithm based on BERT shows promising results in predicting suicide risk from textual data. Its superior performance, compared to traditional models and even some advanced deep learning models, highlights its potential for such applications. The BERT model's ability to understand context and semantic nuances in text data provides a significant advantage. Future work could involve expanding the dataset, exploring additional preprocessing techniques, and experimenting with more advanced neural network architectures to further improve prediction performance. Additionally, the integration of multimodal data (e.g., combining text with social media activity patterns) could offer more comprehensive insights into suicide risk prediction. Building upon the promising results of this study, several avenues for future research can be explored to further enhance suicide risk prediction models:

- *Domain-Specific Fine-Tuning:* Additional fine-tuning of BERT on domain-specific text data, such as mental health forums, social media posts related to mental health, and clinical notes, to improve the model's understanding and predictive capabilities in the context of suicide risk.
- *Evaluating Other Transformer-Based Models:* Assessment of other advanced transformer-based models like RoBERTa, XLNet, and GPT-3 to compare their performance against BERT in predicting suicide risk. This comparison can provide insights into the strengths and weaknesses of different architectures for this specific task.
- *Expanding the Dataset:* Incorporating larger and more diverse datasets from various sources, including clinical records, social media, and other online platforms, to enhance the robustness and generalizability of the model.
- *Exploring Additional Preprocessing Techniques:* Investigating more sophisticated text preprocessing techniques, such as lemmatization, stemming, and the removal of specific types of noise, to see their impact on model performance.
- *Multimodal Data Integration:* Combining textual data with other data modalities, such as audio, video, and physiological signals, to create a more holistic and accurate prediction model. This can provide a more comprehensive understanding of an individual's mental state.
- *Experimenting with Advanced Neural Network Architectures:* Implementing and testing more complex neural network architectures, such as attention mechanisms, graph neural networks, and hybrid models that combine convolutional and recurrent layers, to further improve predictive performance.
- *Developing Real-Time Prediction Systems:* Creating real-time prediction systems that can be integrated into online platforms for early detection and intervention. This can potentially save lives by providing timely support and resources to individuals at risk.
- *Addressing Ethical and Privacy Concerns:* Ensuring that the deployment of these models respects user privacy and adheres to ethical guidelines. Developing frameworks for data anonymization, secure data handling, and responsible AI usage is crucial.
- *Bias in Social Media Data:* Social media data can introduce biases based on demographic factors, such as age, gender, and cultural background. In this study, we acknowledge that the dataset may not fully represent the global population. To mitigate bias, future work will incorporate demographic balancing techniques and explore dataset diversity to ensure that the model performs fairly across different groups.
- *Model Interpretability for Clinician:* In suicide prevention, it is critical that clinicians can understand and trust the predictions made by AI models. While BERT excels in accuracy, it is often criticized for being a "black box." To address this, future research will be exploring attention-based mechanisms and SHAP (Shapley Additive Explanations) to provide insights into why the model makes certain predictions. This will allow clinicians to better interpret and act on the model’s outputs.

The potential for integrating this BERT-based model into clinical practice is promising, but more work is needed for validation in real-world settings. Future research will involve pilot studies in collaboration with mental health professionals to assess the model’s effectiveness in clinical environments. Additionally, user feedback will guide the development of explainable AI tools to support clinical decision-making.

These future research directions aim to build on the current study's findings, improving the effectiveness and reliability of suicide risk prediction models. By continuously refining and expanding upon these models, the goal is to create tools that can aid in early detection and intervention, potentially saving lives and improving mental health outcomes.

References

[1] Martínez-Castaño, R., Htait, A., Azzopardi, L., & Moshfeghi, Y. (2021). BERT-based transformers for early detection of mental health illnesses. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12 (pp. 189-200). Springer International Publishing.

[2] Grimland, M., Benatov, J., Yeshayahu, H., Izmaylov, D., Segal, A., Gal, K., & Levi-Belz, Y. (2024). Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box. Suicide and Life-Threatening Behavior.

[3] Jacobucci, R., Ammerman, B. A., & Tyler Wilcox, K. (2021). The use of text-based responses to improve our understanding and prediction of suicide risk. Suicide and Life-Threatening Behavior, 51(1), 55-64.

[4] Fonseka, T. M., Bhat, V., & Kennedy, S. H. (2019). The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. Australian & New Zealand Journal of Psychiatry, 53(10), 954-964.

[5] Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., ... & Yaniv-Rosenfeld, A. (2021). Big Data analytics and artificial intelligence in mental healthcare. In Applications of big data in healthcare (pp. 137-171). Academic Press.

[6] Sweeney, C., Ennis, E., Mulvenna, M. D., Bond, R., & O'Neill, S. (2024). Insights derived from text-based digital media, in relation to mental health and suicide prevention, using data analysis and machine learning: systematic review. JMIR mental health, 11, e55747.

[7] Burkhardt, H. A., Ding, X., Kerbrat, A., Comtois, K. A., & Cohen, T. (2023). From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction. Journal of the American Medical Informatics Association, 30(6), 1068-1078.

[8] Ophir, Y., Tikochinski, R., Asterhan, C. S., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. Scientific reports, 10(1), 16685.

[9] Malgaroli, M., Hull, T. D., Bantilan, N., Ray, B., & Simon, N. (2020). Suicide Risk Automated Detection Using Computational Linguistic Markers from Patients' Communication With Therapists. Biological psychiatry, 87(9), S444.

[10] Levis, M., Levy, J., Dufort, V., Gobbel, G. T., Watts, B. V., & Shiner, B. (2022). Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry research, 315, 114703.

[11] Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., ... & McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. PloS one, 9(1), e85733.

[12] Naseem, U., Khushi, M., Kim, J., & Dunn, A. G. (2022). Hybrid text representation for explainable suicide risk identification on social media. IEEE transactions on computational social systems.

[13] Baghdadi, N. A., Malki, A., Balaha, H. M., AbdulAzeem, Y., Badawy, M., & Elhosseini, M. (2022). An optimized deep learning approach for suicide detection through Arabic tweets. PeerJ Computer Science, 8, e1070.

[14] Cohen, J., Wright-Berryman, J., Rohlfs, L., Trocinski, D., Daniel, L., & Klatt, T. W. (2022). Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. Frontiers in digital health, 4, 818705.

[15] Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science, 5(3), 457-469.

[16] Levis, M., Westgate, C. L., Gui, J., Watts, B. V., & Shiner, B. (2021). Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychological medicine, 51(8), 1382-1391.

[17] Barak-Corren, Y., Castro, V. M., Nock, M. K., Mandl, K. D., Madsen, E. M., Seiger, A., ... & Smoller, J. W. (2020). Validation of an electronic health record–based suicide risk prediction modeling approach across multiple health care systems. JAMA network open, 3(3), e201262-e201262.

[18] Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, 34(13), 10309-10319.

[19] Kumari, M., Singh, G., & Pande, S. D. (2024). Depressonify: BERT a deep learning approach of detection of depression. EAI Endorsed Transactions on Pervasive Health and Technology, 10.

[20] Sawhney, R., Joshi, H., Gandhi, S., Jin, D., & Shah, R. R. (2021). Robust suicide risk assessment on social media via deep adversarial learning. Journal of the American Medical Informatics Association, 28(7), 1497-1506.

[21] McCoy Jr, T. H., Pellegrini, A. M., & Perlis, R. H. (2019). Research Domain Criteria scores estimated through natural language processing are associated with risk for suicide and accidental death. Depression and anxiety, 36(5), 392-399.

[22] Arowosegbe, A., & Oyelade, T. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. International Journal of Environmental Research and Public Health, 20(2), 1514.

[23] Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., ... & Ling, X. B. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Translational psychiatry, 10(1), 72.

[24] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), 7.

[25] Tsui, F. R., Shi, L., Ruiz, V., Ryan, N. D., Biernesser, C., Iyengar, S., ... & Brent, D. A. (2021). Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA open, 4(1), ooab011.

[26] Velupillai, S., Hadlaczky, G., Baca-Garcia, E., Gorrell, G. M., Werbeloff, N., Nguyen, D., ... & Dutta, R. (2019). Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Frontiers in psychiatry, 10, 36.

[27] Moradian, H., Lau, M. A., Miki, A., Klonsky, E. D., & Chapman, A. L. (2023). Identifying suicide ideation in mental health application posts: A random forest algorithm. Death Studies, 47(9), 1044-1052.

[28] Bayramli, I., Castro, V., Barak-Corren, Y., Madsen, E. M., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2022). Temporally informed random forests for suicide risk prediction. Journal of the American Medical Informatics Association, 29(1), 62-71.

[29] Howard, D., Maslej, M. M., Lee, J., Ritchie, J., Woollard, G., & French, L. (2020). Transfer learning for risk classification of social media posts: model evaluation study. Journal of medical Internet research, 22(5), e15371.

[30] Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. Journal of affective disorders, 245, 869-884.
[31] Suicide and Depression Detection Dataset: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch/data Accessed on :15/06/2024.

Authors' Profiles

Dr. Karan Bajaj, affiliated with Lovely Professional University, focuses on areas like Edge/Fog Computing, IoT, and Machine Learning. His research includes topics such as IoT-based offloading frameworks, WSN integration with IoT, and intrusion detection systems. His most cited work is the "Implementation analysis of IoT-based offloading frameworks on cloud/edge computing," with 112 citations as of 2022. Bajaj has contributed significantly to the field, particularly in integrating wireless sensor networks with IoT and analyzing machine learning models' performance in big data contexts.

Dr. Mukesh Kumar is an Associate Professor at Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India. He holds a Ph.D. in Computer Science from Himachal Pradesh University, Shimla, where he specialized in designing ensemble and hybridized data mining techniques for analyzing academic performance. With over 15 years of experience in academia, Dr. Kumar has held various teaching positions, including Assistant Professor at Lovely Professional University and Chitkara University. He has contributed to various international conferences and holds intellectual property rights for several innovations. He is also a reviewer for multiple prestigious journals and a member of several professional organizations, including the International Association of Engineers and the Indian Society for Technical Education.

Dr. Shaily is PhD in Computer Science in the year 2014. Prior to this she had her Masters in technology in 2009. She has a total of 18 years of experience in teaching. Her research interests include multiprocessor systems on chip, embedded systems, wireless networks, security in networks, data mining and Education Engineering. She is student branch coordinator and member in management of Computer society of India (CSI). She is senior member of IEEE and member of ACM and ISTE. She has in total 42 SCI, reputed international journals and conference papers in her credits. She has filed 10 patents also. She has guided 10 M.Tech research students and 6 PhD students. She is also working as Assistant Faculty in IUCEE IIEECP program in India. She has been recognized for her contribution in Teaching Learning by IUCEE and an IGIP certified faculty. Her current affiliation is Professor and Dean (CSE-AE) in Chitkara University, Punjab.

Dr. Vivek Bhardwaj, PhD, is an accomplished academic and researcher currently serving as an Assistant Professor in the Department of Computer Science and Engineering at Manipal University Jaipur, Rajasthan. His expertise spans over eight years of teaching and research, during which he has made significant contributions to his field. Dr Bhardwaj's extensive experience in academia is complemented by his role as a certified Robotic Process Automation trainer. RPA is a technology that automates repetitive tasks using software robots, improving efficiency and accuracy in various processes. His certification in this domain highlights his proficiency in both the theoretical and practical aspects of RPA, making him a valuable resource for students and professionals looking to enhance their skills in this cutting-edge technology.

Sahil Walia is a Data Solutions Architect at Snowflake, Inc., located in Georgia, USA. He works within the Professional Services division, assisting Snowflake's customers in optimizing, accelerating, and achieving their business objectives. Sahil holds a Master's degree in Information Management from Syracuse University. With over a decade of industry experience, he has successfully led numerous data migration and modernization projects that have resulted in cost-efficiency, improved performance, and streamlined data pipelines. Sahil is actively involved in the research community and has been a dedicated member of the IEEE.

How to cite this paper: Karan Bajaj, Mukesh Kumar, Shaily Jain, Vivek Bhardwaj, Sahil Walia, "Enhancing Suicide Risk Prediction through BERT: Leveraging Textual Biomarkers for Early Detection", International Journal of Intelligent Systems and Applications(IJISA), Vol.17, No.2, pp.101-111, 2025. DOI:10.5815/ijisa.2025.02.06