Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

Enhancing AI Depression Detection Using Transfer Learning

Na Wang[1,2,3*], Raja Kamil[1], Syed Abdul Rahman Al-Haddad[1], Normala Ibrahim[4], Zhen Zhao[5]

[1] Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, Selangor, 43400, Malaysia
[2] School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW, 2308, Australia
[3] School of Automation, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, 510665, China
[4] Faculty of Medicine and Health Sciences/Hospital Sultan Abdul Aziz Shah, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia
[5] Department of Electrical Engineering, Faculty of Engineering, Universiti Putra Malaysia, Kuala Lumpur, Malaysia
E-mail: na.wang10@uon.edu.au

Received: 3 December 2024; **Revised:** 7 January 2025; **Accepted:** 10 January 2025

Abstract: Depression is a serious mental health disorder that poses significant challenges to individuals' emotional well-being, daily functioning, and overall quality of life. While artificial intelligence methods offer promising solutions in diagnosing, their development typically requires large datasets, which are difficult due to privacy concerns and the complex nature of clinical diagnosis. To address these challenges, this study leverages transfer learning to improve depression detection. Specifically, the method employs the bidirectional encoder representations from transformers (BERT) pre-trained model on large-scale language data, and fine-tunes it on the Extended Distress Analysis Interview Corpus to adapt the model for depression detection. By using the BERT Tokenizer, interview transcripts were tokenized to retain critical linguistic context. These tokens were then processed by the pre-trained bert-base-uncased model to extract robust language features. The features were then passed through a 1-dimensional convolutional neural network (1DCNN) for further analysis, enabling the detection of depression-related patterns. Finally, the refined features were classified via dense layers. To investigate the effectiveness of this method, a total of 11 models-six conventional machine learning models and five neural networks-were tested using three tokenization methods for comparison. Among them, the BERT Tokenizer + 1DCNN model achieved the best performance, with an accuracy of 89.3%, F1 score of 89.4%, and AUC of 95.0%. Notably, transfer learning improved accuracy by 7.7%, highlighting its effectiveness in training neural networks on small datasets. These results demonstrate that the proposed approach not only addresses the issue of limited training data but also significantly enhances the accuracy and reliability of AI-aided depression detection. The method can be applied in computer-aided systems for improving clinical depression diagnostics and other healthcare applications where data scarcity is a barrier.

Keywords: depression, text, BERT, transfer learning, 1DCNN, extended distress analysis interview corpus (E-DAIC)

MSC: 62M45, 62P15

Abbreviation

1DCNN	1-dimensional convolutional neural network
AI	Artificial Intelligence

DOI: https://doi.org/10.37256/cm.6320256184

1. Introduction

Depression, also known as major depressive disorder (MDD), is a prevalent psychological illness that affects a significant portion of the global population. According to the World Health Organization (WHO), around 280 million people in the world suffer from depression. This disorder entails many symptoms, including loss of interest, anger, pessimism, changes in weight, feelings of worthlessness, and at its worst could lead to suicide.

The depression diagnosis process is complicated and requires expert knowledge of a professional psychiatrist. Early detection plays an important role in diagnosing and effective treatment. But it is not easy due to many reasons. Mental health services can be expensive, and this economic barrier can prevent people from seeking help. In many regions, particularly rural or remote areas, healthcare facilities equipped to diagnose and treat depression are scarce or even nonexistent [1]. Artificial intelligence (AI) technology has developed rapidly in recent years and has been applied in a variety of applications including depression diagnosis.

Although AI-driven automatic mental health services promise to bridge some gaps, AI models for depression detection face significant challenges due to insufficient training data. Collecting clinically diagnosed depression data is particularly difficult due to privacy concerns, ethical considerations, and limited medical resources. Training and testing AI methods on data from clinically diagnosed patients ensures the reliability and validity of the data, providing a solid foundation for developing models tailored to clinical applications. Furthermore, depression, as a mental health illness, often involves hidden behaviors and emotional expressions, with performance and symptoms varying widely from patient to patient. This variability makes it especially challenging to collect sufficient, representative data. Consequently, the translation and adoption of AI technologies in clinical settings for depression healthcare often struggle due to the difficulties of generalizing systems trained on small datasets to real-world patient populations.

During the depression diagnosis process, a psychiatrist organizes interviews with the patient, and diagnoses based on standardized assessment tools such as the beck depression inventory (BDI) [2], self-rating depression scale (SDS) [3], Hamilton rating scale for depression (HAMD, sometimes also abbreviated as HRSD) [4], the major depression inventory (MDI) [5], and the patient health questionnaire (PHQ) [6].

Most assessment tools are based on interviews and questionnaires, which indicates that what the patient says (text content of the speech) provides valuable information for the detection of depression. Textual features (sentiment markers) refer to signature words and phrases that indicate depressive emotion. For example, absolutist words have been established as semantic markers for depression (e.g., NEVER in "I NEVER want to wake up") [7], as well as phrases such as "I am so tired all the time". Compared to other modalities (e.g. facial expression and voice characteristics), textual features are straightforward and have the advantage of explainability. A study conducted by Liu et al. [8] examined 219 participants according to the density of their text message data and found significant correlations between lexica and mean PHQ-8.

When engineering for automatic depression detection, the required text data are still scarce. The rise of artificial intelligence (AI) offers a promising avenue for addressing these challenges, particularly through automated and objective methods for depression detection. However, developing reliable AI systems is complicated by the scarcity of high-quality, clinically labeled training data and the nuanced, context-dependent nature of language used by individuals with depression. Existing models often struggle to generalize across diverse populations and languages, limiting their clinical applicability.

This study seeks to address these challenges by leveraging transfer learning with BERT, a state-of-the-art language model pre-trained on massive text corpora. By combining BERT's ability to extract rich contextual embeddings with advanced neural architectures, we aim to develop a robust system for detecting depression from text data. The research is motivated by the need to enhance the accuracy and reliability of AI-based tools in real-world clinical contexts, ultimately contributing to more accessible and effective mental health care solutions.

Specifically, we propose a model that integrates a pre-trained large language model (LLM) with a 1-dimensional convolutional neural network (1DCNN) for automatic depression detection. The proposed method utilizes training data annotated based on clinical diagnoses, leveraging the pre-trained LLM's deep understanding of language patterns to enhance the model's capacity for identifying depression. This approach not only addresses the scarcity of clinically labeled training data but also improves the model's generalizability and effectiveness in detecting depression across diverse contexts.

The contributions of this work can be summarized as follows:

1. Hybrid Model Design: We propose a hybrid model that combines a large language model (LLM) with a 1DCNN deep learning architecture for effective depression detection.

2. Addressing the Small Dataset Challenge: By utilizing the knowledge of a pre-trained large language model, we enhance the detection efficiency of the deep learning model, even when trained on a small dataset, thereby addressing the challenge of insufficient training data in depression detection tasks.

3. Ablation Experiments: We conduct comprehensive ablation experiments to verify the effectiveness of transfer learning techniques in adapting LLMs for specialized medical tasks.

4. Reference for Researchers: Through the proposed transfer learning method and comparative experiments, we provide valuable insights and a reference framework for researchers in the field.

The rest of this paper is organised as follows. Section 2 introduces the previous research related to this study; Section 3 introduces the dataset used in this study, explains the data processing, the model structure, and the workflow; Section 4 introduces the experiments setups and the evaluation metrics; Section 5 demonstrates and discusses the results; and finally conclusions are drawn in Section 6.

2. Related research

Text content provides valuable information for depression assessment. Many researchers have tried to develop automatic depression detection methods using text content. 1DCNN and recurrent neural network (RNN) (and its variants such as gate recurrent unit (GRU), long short-term memory (LSTM) and bidirectional long short-term memory (BiLSTM)) are commonly used deep learning models in these tasks.

CNN was first proposed by LeCun et al.'s early work for image analysis. The extension of CNNs to 1D data (e.g., time-series data, signals) was later adapted in the mid-2010s for tasks such as signal processing, time-series analysis, and natural language processing. Kim et al. reported their study of applying CNNs to text classification in 2014 [9]: this is considered an early effort of 1D CNN applied to word processing. Wang et al. proposed an approach that explicitly applies 1D CNNs to time-series data at the International Conference on Web-Age Information Management held in 2015.

LSTM network was proposed by Hochreiter and Schmidhuber in 1997 [10]. The LSTM architecture was designed to address the vanishing gradient problem in traditional recurrent neural networks (RNNs), enabling the model to learn long-term dependencies more effectively. Based on LSTM, BiLSTM was proposed by Graves and Schmidhuber [11]. This architecture was first introduced for Framewise phoneme classification. The idea was to combine forward and backward LSTM networks to capture dependencies in both directions for sequence processing tasks. LSTM and BiLSTM networks are widely used for a variety of sequence processing tasks such as speech recognition, text classification, machine translation, etc.

In 2022, Wani et al. reported their research on depression screening [12]. The study was conducted using Word2Vec [13] and term frequency-inverse document frequency (TF-IDF) [14] combined with deep learning models: convolutional neural network (CNN) and LSTM. The data was collected from Facebook, Twitter, and YouTube. This study achieved accuracy of 99.02% and 99.01% on the Word2Vec LSTM and Word2Vec (CNN + LSTM) models. The results are very impressive. However, the data for depression recognition were not collected through clinically diagnosed patients.

Transfer learning technology has been applied in a variety of fields, including text-based prediction tasks. This technique has been proven to be effective in many language-related tasks.

In 2018, Kratzwald [15] proposed a tailored form of transfer learning method for affective computing-sent2affect. The network was pre-trained for a sentiment analysis task, and then the output layer was fine-tuned for emotion recognition. The resulting performance confirms both recurrent neural networks (RNNs) and transfer learning outperform traditional machine learning (ML).

In 2022, Choudhry et al. [16] proposed a multitasking framework for fake news and rumour detection, predicting both the emotion and legitimacy of the text. Several deep learning models for single tasks and multitasks were trained.

The performance of the multitasking approach proves the efficacy for better generalization through multi-dataset training and verifies that emotions act as a domain-independent feature.

In 2023, an end-to-end transformer-based multitask framework for sentiment identification and emotion recognition from the SentiMix code-mixed dataset was proposed by Ghosh et al. [17]. In order to maximise the effectiveness of transfer learning and raise the method's overall efficiency, the study employed task-specific data. The study's findings demonstrated how the application of transfer learning to an auxiliary task-emotion recognition-could boost the effectiveness of the main task-sentiment detection.

The pre-trained BERT model [18] was trained through unsupervised learning techniques on the BooksCorpus (800 M words) [19] and English Wikipedia (2,500 M words). The training used a document-level corpus rather than a shuffled sentence-level corpus in order to extract long contiguous sequences. The massive amount of training data ensures the LLM's ability of basic language understanding.

In 2023, Milintsevich et al. [20] reported their study of training a multi-target hierarchical model to predict individual depression. Unlike the aforementioned study, this study used symptoms from patient-psychiatry interview transcripts of the DAIC dataset, in which the patients were clinically diagnosed with depression. The method adopted a model based on RoBERTa (Robustly Optimized BERT) [21], and yielded a 73.9 macro-F1 binary depression estimation score.

In 2024, Zhang et al. [22] propose a multilevel depression status detection approach based on fine-grained prompt learning, called MDSD-FGPL. The approach utilises two kinds of encoders, BERT and T5, to yield two classification results and mean voting algorithm to output a final result, and achieved an f1-score of 0.8276 in the coarse-grained binary classification. BERT was used in this study to generate text embeddings.

In the same year, Mozhdehi et al. [23] conducted a transfer learning experiment utilizing a pre-trained model Emotional BERT on two datasets: Wang and MELD. The study demonstrated the capability of transfer learning techniques to capture the better contextual meaning of the text and for emotion recognition using a small amount of data.

The above studies collectively highlight that transfer learning, particularly with models like BERT, effectively addresses the challenge of limited training data by leveraging pre-trained embeddings and adapting them to specific tasks. Compared to traditional methods, transfer learning demonstrates superior reliability, capturing rich contextual information and generalizing well across tasks and datasets. This makes it particularly advantageous for sensitive applications such as depression detection, where data scarcity and variability pose significant challenges.

However, despite its potential, the application of transfer learning for improving the efficiency of depression detection remains underexplored, especially when faced with limited clinically diagnosed depression data. This study addresses this gap by employing a transfer learning technique, specifically leveraging the pre-trained BERT model, to tackle the issue of insufficient training data in automatic depression detection.

3. Methodology

The objective of this study is to improve text depression detection using transfer learning.

To this end, we leverage the knowledge of a LLM, gained during its extensive pre-training phase. This knowledge includes understanding grammar, general language nuances, and linguistic patterns, which is transferred to the depression detection task. This reduces the need for a large labelled dataset for the new task, making it suitable for scenarios where labelled data is limited. A functional sense of knowledge based transfer learning is demonstrated in Figure 1.

To obtain a model for depression, we designed a hybrid model consisting of BERT base-uncased and 1DCNN. During training, the combined BERT and 1DCNN model is fine-tuned on the specific depression detection dataset. This process leverages the pre-trained model's knowledge of language to increase the model's performance on the depression detection task.

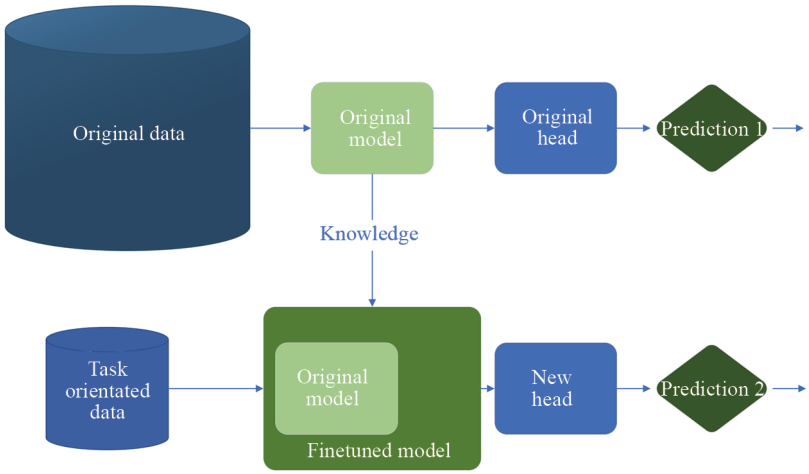


Figure 1. Knowledge-based transfer learning

The method starts with preprocessing the raw text transcripts that were provided by the DAIC dataset, and then feeding them into a deep learning pipeline. The pipeline consists of a BERT model for feature extraction, which includes 12 transformer blocks, 768 hidden units per layer and 12 attention heads (110 M parameters), followed by two BiLSTM layers to capture temporal dependencies, and a Dense layer with a Softmax activation for the final classification. This approach leverages the strengths of both transformer-based models and recurrent neural networks to achieve higher classification accuracy on a small dataset. See Figure 2 for the workflow of the proposed method.

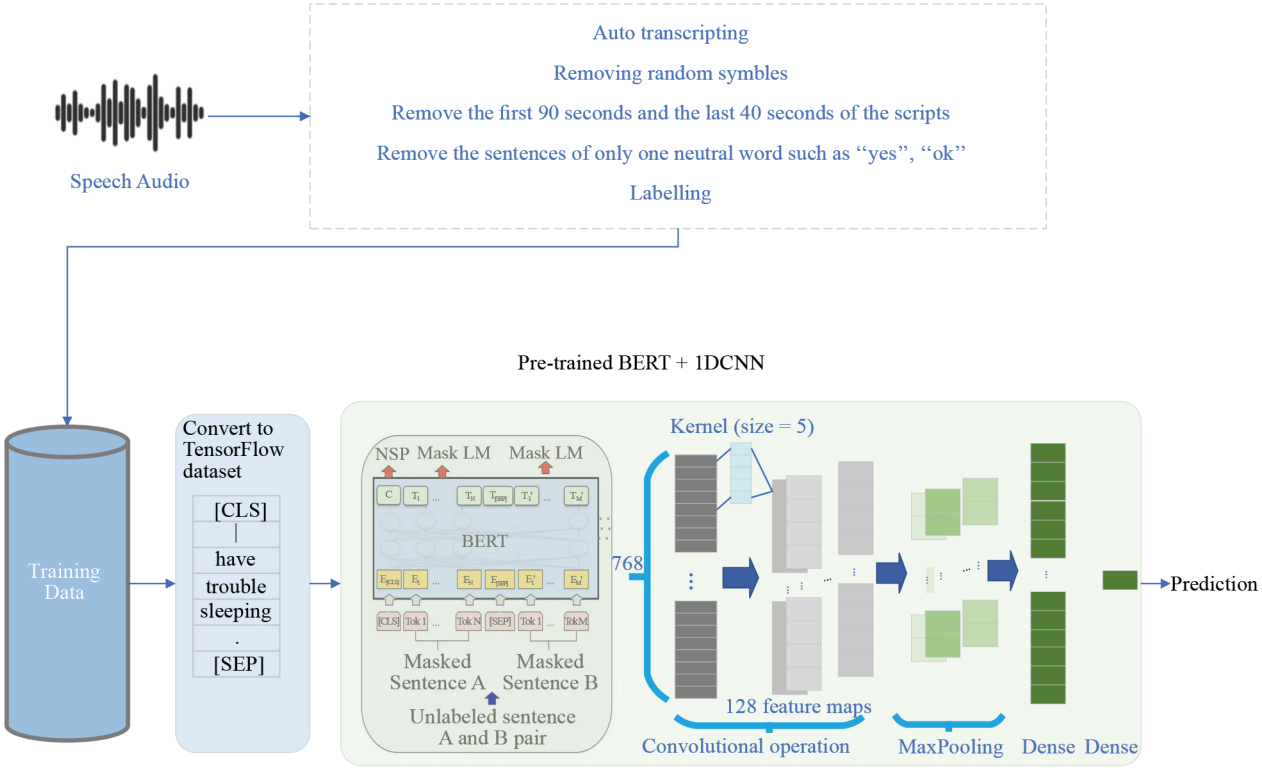


Figure 2. The workflow of the proposed method

3.1 *Dataset and preprocessing*

The tested models including the proposed model were trained on the E-DAIC dataset, which was created based on clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD). The E-DAIC dataset provides audio recordings of the conversations between an interviewer and a participant; textual transcripts of the audio recordings, including both the interviewer's and the participant's spoken words; and metadata-information about the participants, such as gender, and psychological assessment scores. Video recordings are not available, but the video features extracted by different algorithms are provided.

A subset of the sessions was collected where the virtual agent is controlled by a human interviewer (wizard) in another room. The other subset was collected using an AI-controlled agent, who acts in a fully autonomous way using different automated perception and behaviour generation modules. The interviewees were veterans of the U.S. military and members of the general public from the Los Angeles area. There are 275 participants recorded in this dataset. According to the PHQ-8 binary labels, 66 participants (35 male, 31 female) were suffering from MDD at the time of recording, and 209 (135 male, and 74 female) were HCs. Table 1 and Table 2 summarise the dataset information.

Table 1. Data collection methods of E-DAIC dataset

Psychiatric disorder	Diagnosis method	No. of participants	Data modality
Depression	PHQ-8	275	Visual, Audio, Text
PTSD	PCL-C[2]		

[2] We acknowledge that even though the participants were evaluated with PHQ-8 and PCL-C, the labels in this study adopted were PHQ-8 binary scores

Table 2. Number of participants and the positive ratio of the E-DAIC dataset

Gender	MDD	HC	Positive ration
Female	31	74	29.5%
Male	35	135	20.6%
Total	66	209	24%

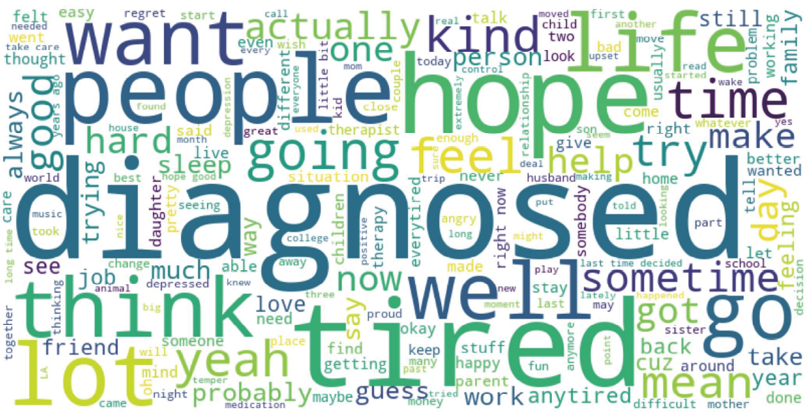


Figure 3. Word cloud of the MDD group text script

When processing the text script of the depressed participants, a pattern of negativity that reflects the feeling of hopelessness and tiresomeness was observed. A word cloud visual demonstration of the MDD text scripts is given in Figure 3.

To prepare the data for training, we 1) removed the first 90 seconds and the last 40 seconds of the scripts, because these are the parts of the introductions to the interviewee about how the interview would be conducted, general questions such as "where are you from", greetings and goodbyes of the conversations; 2) removed the sentence that contains only one neutral word such as "yes", "ok"; 3) labelled the scripts based on the PHQ-8 binary scores, "1" denotes MDD and "0" denotes HC; 4) randomly removed 50% of the HCs to mitigate the imbalance problem between MDDs and HCs, we acknowledge that the removing causes the reduction of the training data's size, but it generates a more balanced data as our primary concern is to detect the MDD class. Finally, these labelled data were saved as CSV files. See Table 3 for a preview of the training data.

Table 3. Preview of the training data

Text	Label
I try I'm trying I'm trying I'm trying I'm trying I'm trying I'm trying	1
It can be tough to find a good job these days I don't even care about it	1
I survived day by day	1
I applied from anywhere and everywhere	1
I just I don't know if I have what it takes to continue to do	1
Learning about different cultures	0
I just recently went to the United Kingdom for a wedding 2 years ago or last year	0
I have a very close relationship with my family	0
Just recently I went to Italy we went and saw Pisa and Rome and Florence and Natalie	0
I don't feel guilty about anything	0

3.2 *BERT pre-trained model*

In this section, we introduce the BERT pre-trained model for the depression recognition task. The workflow of the model is demonstrated in Figure 4.

Bidirectional encoder representations from transformers (BERT) [24] is a pre-trained model developed by Google that was designed for artificial intelligence to understand the context of words in a sentence by analysing both the preceding and succeeding words.

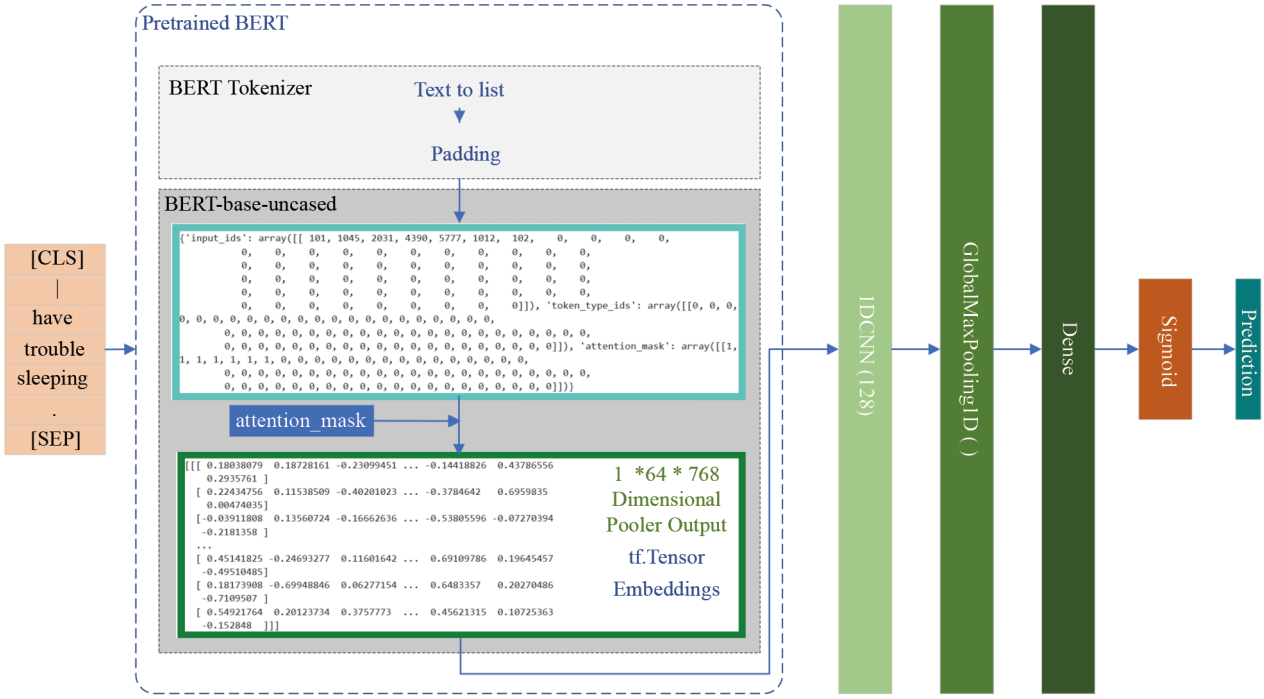


Figure 4. The proposed model

Traditional transformer models [18] usually predict the next word based on the previous inputs. BERT, however, leverages the self-attention mechanism to process tokens in both directions (left to right and right to left). It uses the masked language modelling (MLM) to train the model to predict the masked position from both directions, allowing it to capture context from both ends of a sentence simultaneously. To process the input text, BERT first tokenizes the input text and converts it into a vector using an embedding layer. In addition, it also include position information of the tokens in the sequence and add the positional encodings to the input overall embeddings. The positional encodings for each position *pos* and dimension are defined as:

$$\mathrm{PE}_{(pos,\,2i)} = \sin\left(\frac{pos}{10{,}000^{2i/d_{model}}}\right), \tag{1}$$

$$\mathrm{PE}_{(pos,\,2i+1)} = \cos\left(\frac{pos}{10{,}000^{2i/d_{model}}}\right), \tag{2}$$

where d_{model} is the dimension of the model.

After the text is converted into vectors, the self-attention mechanism is used to allow each token to build a contextual representation based on the importance of other tokens to it. For a given a sequence of input vectors $X = [x_1, x_2, \ldots, x_n]$, self-attention is computed as:

$$Q = XW^{Q}, \tag{3}$$

$$K = XW^K, \tag{4}$$

$$V = XW^V, \tag{5}$$

where Q, K, V are weight matrices for queries, keys, and values. Q represents the current token we are focusing on. For each token in the input sequence, the Q-query matrix represents the "query" vector of the current word (position), which is the focus used by the model to search for relevant information or context in the input sequence, to determine what information needs to be obtained from other locations or words in order to process the current location or word. The self-attention scores are computed by taking the dot product of the Q vector with all K vectors to calculate the relevance score between the current position and other positions in the sequence, as shown below.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

where d_k is the dimension of the key vectors.

BERT uses multi-head self-attention to allow the model to focus on different parts of the sentence simultaneously. Multiple attention heads are computed in parallel and then concatenated:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O, \tag{7}$$

where W^O is a weight matrix used to linearly transform the concatenated outputs of multiple attention heads, each head_i is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

Each position in the sequence is passed through a feed-forward neural network independently and identically.

$$f(x) = \text{ReLU}(0, xW_1 + b_1)W_2 + b_2, \tag{9}$$

where W_1 and W_2 are the weight matrices, and b_1 and b_2 are biases.

Residual connections x'' and layer normalization x' are as follows.

$$x' = \text{LayerNorm}(x + \text{MultiHead}(Q, K, V)) \tag{10}$$

$$x'' = \text{LayerNorm}(x' + f(x')) \tag{11}$$

The training strategy of BERT-MLM randomly masks tokens in the input and predicts them based on the surrounding context. The objective of the MLM is to predict the original tokens given randomly masked tokens within a sequence. Given a sequence of tokens $(w_1, w_2, ..., w_n)$, randomly mask a percentage of tokens, and replace masked tokens with a

special token '[MASK]'. The model predicts the original tokens using the probability P distribution over the vocabulary. as shown below:

$$\mathscr{L}_{\text{MLM}} = -\sum_{i} \log P(w_i' \mid w_{\text{context}}). \tag{12}$$

The NSP objective predicts if a pair of sentences are consecutive in the original text. Given a pair of sentences (S_A, S_B), predict if S_B follows S_A:

$$\mathscr{L}_{\text{NSP}} = -\log P(\text{IsNext}) - \log P(\text{NotNext}). \tag{13}$$

3.3 *1 dimensional convolutional neural network (1DCNN)*

1DCNNs are suitable for sequential data such as the texts in this task. The architecture of an 1DCNN is similar to the more commonly known 2DCNNs used for image processing, only that they are applied for one dimension data. Given an input sequence $x = (x_1, x_2, ..., x_n)$, the 1 dimensional convolution operation applies a filter $w = (w_1, w_2, ..., w_m)$ on the input x using the following equation:

$$y_i = \sum_{j=1}^{m} x_{i+j-1} \cdot w_j, \tag{14}$$

where y_i is the result at the ith position.

The Pseudocode of the proposed method is demonstrated in Algorithm 1.

Algorithm 1 Pseudocode of the proposed method

Input: Raw text data $\mathbf{X}_{text} = \{x_1, x_2, \ldots, x_N\}$ and labels $\mathbf{y}$, where N is the number of samples

Output: Predictions $\mathbf{y}_{pred}$ for binary classification

Load dataset and extract the 'text' and 'label' columns

Split dataset into training and test sets using `train_test_split`

Step 1: Encode labels

Encode target labels $\mathbf{y}$ using `LabelEncoder`

Step 2: Load pre-trained BERT model and tokenizer

Load `BertTokenizer` and `BertModel` from Hugging Face

Step 3: Extract BERT embeddings

for each text sample x_i in $\mathbf{X}_{train}$ and $\mathbf{X}_{test}$ do

  Tokenize x_i with BERT tokenizer

  Pass tokenized input to BERT model to get hidden state embeddings

  Compute the mean of embeddings across the sequence length

end for

Step 4: Scale the BERT embeddings

Apply `StandardScaler` to training and test embeddings

Step 5: Initialize 1D CNN model

Create a `Sequential` model with the following layers:

• Conv1D with 128 filters, kernel size 3, ReLU activation

• GlobalMaxPooling1D layer

• Dropout layer with rate 0.5

• Dense layer with 256 units and ReLU activation
• Dropout layer with rate 0.5
• Dense output layer with 1 unit for binary classification (sigmoid activation)

Step 6: Compile the 1D CNN model

Compile the model using Adam optimizer, binary cross-entropy loss, and accuracy as the metrics.

Step 7: Train the model

Train the model using the scaled BERT embeddings for 30 epochs with batch size 32, using early stopping

Step 8: Model evaluation

Evaluate the model on test set embeddings and output the predictions $\mathbf{y}_{pred}$.

4. Evaluation metrics and experiment setups

The experiments were conducted on a workstation of AMD Ryzen 9 CPU, 64 RAM, NVIDIA GeForce RTX 4070 8G GPU; and Windows 11 OS, Jupyter, Python 3.11 programming environment. The objective of this study is to investigate the effectiveness of leveraging transfer learning technology to train, fine-tune and optimize a pre-trained large language model for the special task of depression recognition based on text contents. The experiment results are measured by the following metrics, accuracy, precision, recall, and F1, the equations of these metrics are given below.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{15}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{16}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{17}$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \tag{18}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative counts of the predictions, respectively.

In addition, the receiver operating characteristic (ROC) curves of the learning models are also plotted for comparison. The ROC curve is the plot of the true positive rate against the false positive rate at each threshold setting. In each ROC plot, the area under curve (AUC) was calculated too. A higher AUC indicates better overall performance.

5. Experimental results and discussion

In this section, we investigate the performance of various ML models for text depression detection. Three types of vector-generating methods: 1) Term frequency-inverse document frequency (TF-IDF), which is commonly adopted with traditional machine learning models [25], 2) Keras frequency-based tokenization, which is commonly adopted with neural network models [26], and the pre-trained BERT-base-uncased model were tested under comparison.

5.1 *Experiments of none transfer learning methods*

5.1.1 *TF-IDF tokens*

TF-IDF treats each word as a separate feature. The importance of a word is determined by its TF-IDF score. For a term t in a document D, the TF score is calculated to measure how frequently this term appears in the document, using the following equation:

$$\mathrm{TF}(t, d) = \frac{O_t}{N_d}, \tag{19}$$

where t denotes the term (word), d denotes the document, O_t is the number of occurrences of t in d, and N_d denotes the total number of words in d.

The IDF score measures how important a term is across the entire corpus D:

$$\mathrm{IDF}(t, D) = \log\left(\frac{N}{1+\mathrm{DF}(t)}\right), \tag{20}$$

where N is the total number of documents in the corpus. $\mathrm{DF}(t)$ is the number of documents containing the term t. The TF-IDF score is the product of the Term Frequency and Inverse Document Frequency:

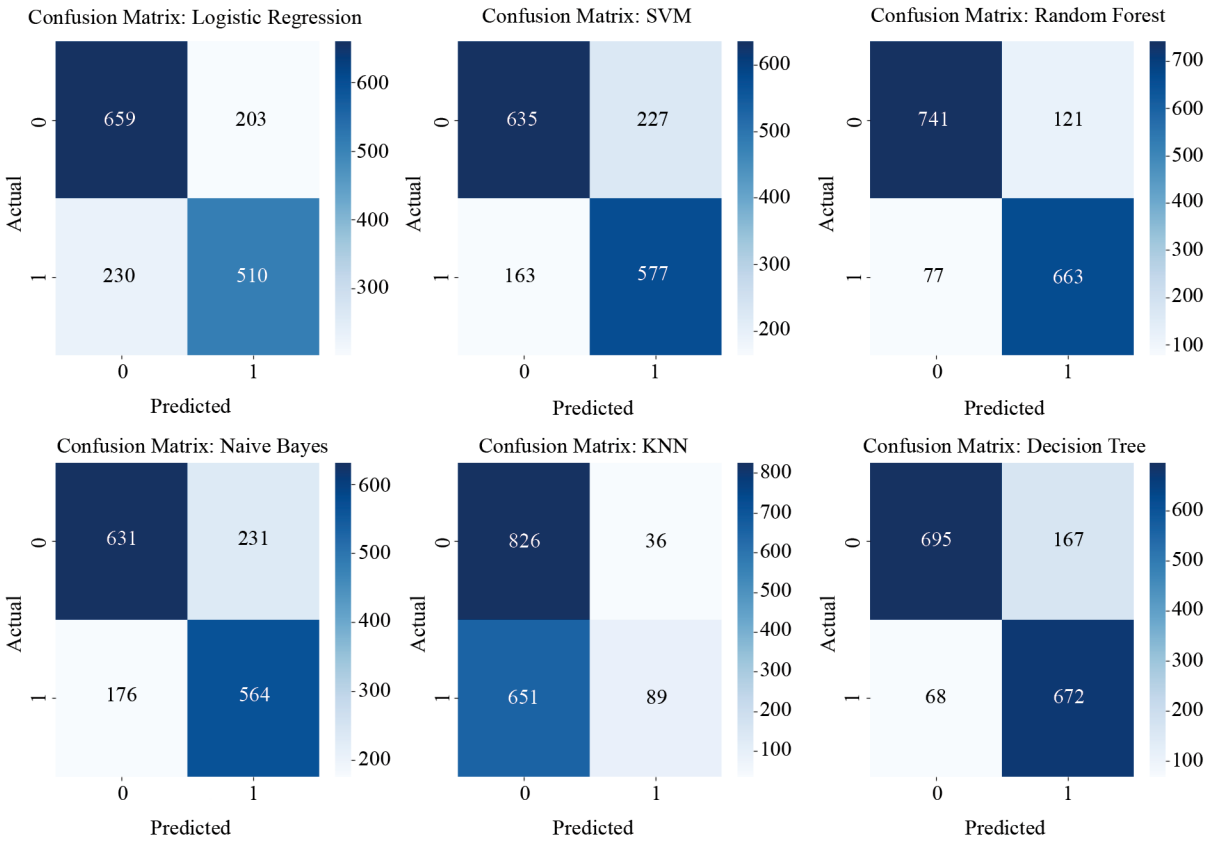
$$\text{TF-IDF}(t, d, D) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t, D). \tag{21}$$

Six conventional ML models, namely logistic regression (LR), support vector machine (SVM), random forest (RF), Naive Bayes (MNB), K neighbours (KNN), and decision tree (DT) were tested using the TF-IDF tokens extracted from the text transcripts. The parameter setting of these models is shown in Table 4.

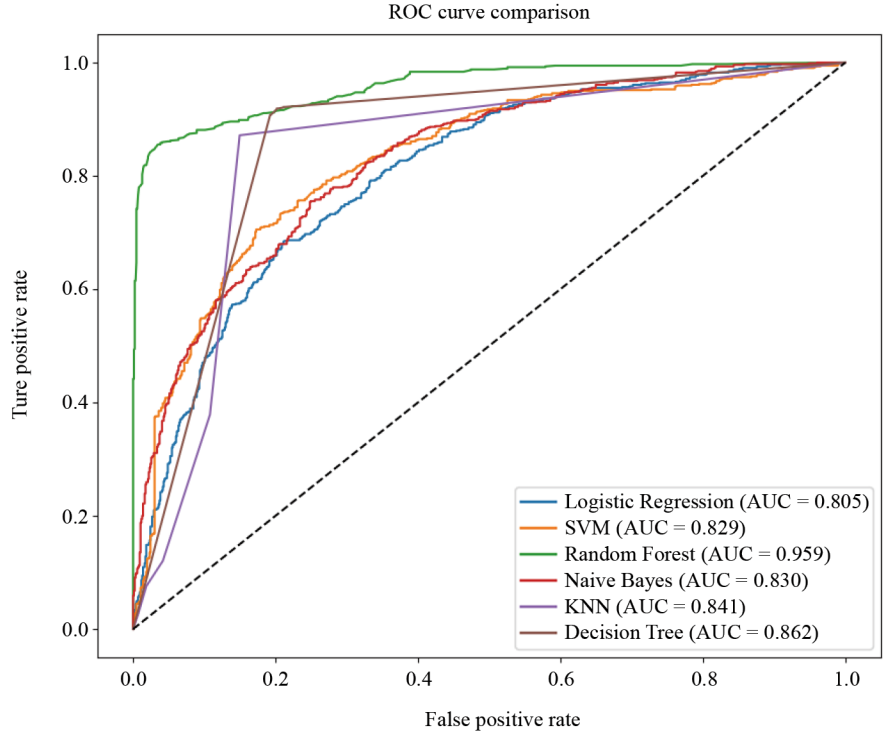
Table 4. The parameter settings of the conventional ML models

No.	Model	Parameters
1	LR	Penalty: L2, optimizer: L-BFGS, max iterations: 100
2	SVM	Kernel: linear
3	RF	Number of estimators: 100
4	NB	Multinomial, alpha: 1.0
5	KNN	Number of neighbors: 5
6	DT	Criterion: ‘gini’

Figure 5 (a) demonstrates the confusion matrices of the conventional ML models. In the confusion matrices, a darker colour indicates a higher number of predictions. As can be seen, the RF, SVM and DT models perform well in terms of balance as the colours on the diagonal of a matrix are darker. The KNN did not perform well in this task as the darker-coloured cubes are on the left side instead of the diagonal.



(a) Confusion matrices



(b) ROC curves

Figure 5. Visual comparison of the conventional ML models using TF-IFD tokens

The evaluation metrics of the conventional ML models using TF-IFD tokens are reported in Table 5. RF model performed well compared to the other models and achieved an accuracy of 87.6%. DT has the second-highest accuracy of 85.3%. The model with the lowest accuracy was KNN, 57.1%.

Figure 5 (b) demonstrates the ROC curves of the conventional ML models. RF reaches a high TPR at a very low FPR, yielding the highest AUC of all the compared models. DT and KNN have higher AUCs than the rest of the models, but KNN's TPR in the area of FPR < 0.1 was the lowest of all.

In the experiment of training neural network models using TF-IDF tokens for depression detection. Five neural networks (NNs), namely RNN, 1DCNN, LSTM, GRU, and BiLSTM were tested under comparison. However, the models did not perform well, suggesting that training neural networks using TF-IDF tokens is not effective for this task.

Table 5. The evaluation metrics of the conventional ML models using TF-IFD tokens

No.	Model	Accuracy (%)	Precision (%)			Recall (%)			F1 (%)			AUC (%)
			0	1	w. avg.	0	1	w. avg.	0	1	w. avg.	
1	LR	73.0	74.1	71.5	72.9	76.5	68.9	73.0	75.3	70.2	72.9	80.5
2	SVM	75.7	79.6	71.8	76.0	73.7	78.0	75.7	76.5	74.7	75.7	82.9
3	RF	**87.6**	90.6	**84.6**	**87.8**	86.0	89.6	**87.6**	**88.2**	**87.0**	**87.7**	**95.9**
4	NB	74.6	78.2	70.9	74.8	73.2	76.2	74.6	75.6	73.5	74.6	83.0
5	KNN	57.1	55.9	71.2	63.0	**95.8**	12.0	57.1	7.6	20.6	47.5	84.1
6	DT	85.3	**91.1**	80.1	86.0	80.6	**90.8**	85.3	85.5	85.1	85.3	86.2

The highest scores of each column are highlighted in **bold**; w. avg.: weighted average

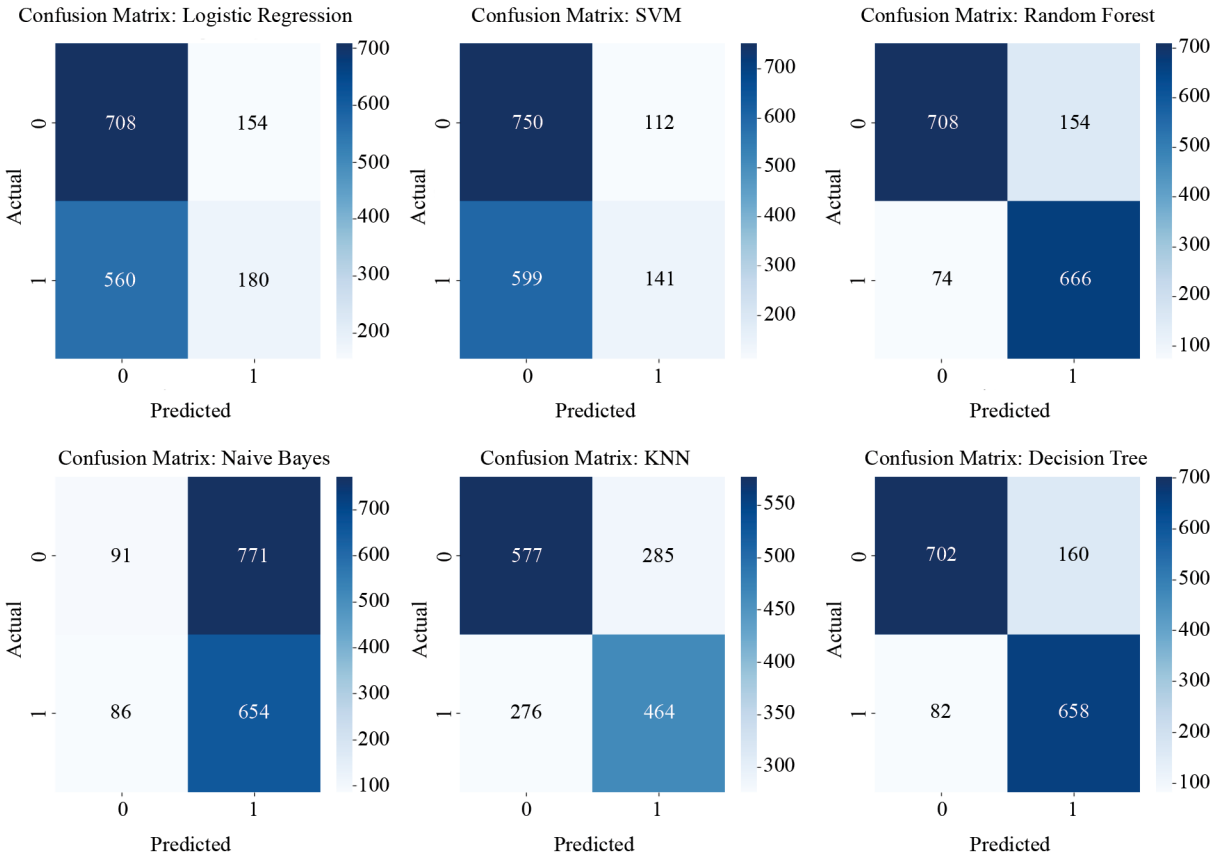
5.1.2 *Keras frequency-based tokens*

In this section, we investigate the performance of the models using the Keras frequency-based tokenization (KFT). The KFT method creates a vocabulary of all unique tokens (words) in the training data. Each unique token in the vocabulary is then assigned a unique integer. The most frequent words are assigned the lowest integers. Let V be the vocabulary size of the corpus, $f(w)$, the words are indexed based on their frequency, index i to each word w_i is assigned based on its frequency rank. For a sentence $S = [w_1, w_2, ..., w_n]$, each word in the sentence is replaced by its corresponding index.

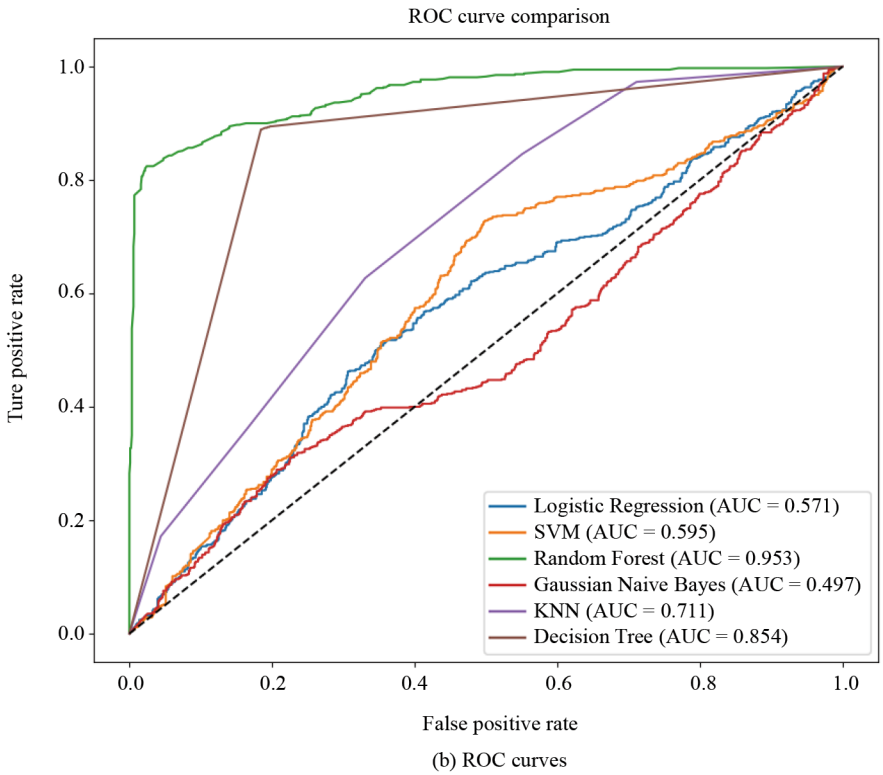
$$T = [index(w_1), indext(w_2), ..., index(w_n)], \tag{22}$$

where T is the tokenized sequence.

Figure 6 (a) and (b) visually demonstrate the classification results through the confusion matrices of these models and the ROC curves. As can be seen, the LR, SVM and NB show poorly balanced performance between the two classes. KNN has a relatively balanced performance between the two classes, but the corrected predicted numbers are low, resulting in an accuracy of 65.0%.



(a) Confusion matrices



(b) ROC curves

Figure 6. Visual comparison of the conventional ML models using KFT tokens

Table 6 demonstrates the accuracy, precision, recall and F1 scores of the conventional ML models using KFT. The parameter settings of these models are the same as Table 4. The experiment results show that the RF outperforms the other models in terms of accuracy, weighted average precision, weighted average recall, and weighted average F1. Following it was the DT model. The LR, SVM and NB did not perform well. SVM has the lowest recall of class 1 (MDD), and NB has the lowest recall of class 0 (HC). In terms of accuracy, all models decreased compared to using TF-IDF tokens, except for KNN.

Table 6. The evaluation metrics of the conventional ML models using KFT tokens

No.	Model	Accuracy (%)	Precision (%)			Recall (%)			F1 (%)			AUC (%)
			0	1	w. avg.	0	1	w. avg.	0	1	w. avg.	
1	LR	55.4 ↓	55.8	53.9	54.9	82.1	24.3	55.4	66.5	33.5	51.3	57.1
2	SVM	55.6 ↓	55.6	55.7	55.7	**87.0**	19.2	55.6	67.8	28.4	49.6	55.6
3	RF	**85.8** ↓	**90.5**	**81.2**	**86.2**	82.1	**90.0**	**85.8**	**86.1**	**85.4**	**85.8**	**95.3**
4	NB	46.5 ↓	51.4	45.9	48.9	10.6	88.4	46.5	17.5	60.4	37.3	49.7
5	KNN	65.0 ↑	67.6	61.9	65.0	66.9	62.7	65.0	67.3	62.3	65.0	71.1
6	DT	84.9 ↓	89.5	80.4	85.3	81.4	88.9	84.9	85.3	84.5	84.9	85.4

↓ indicates decrease compared to using TF-IDF tokens, ↑ indicates increase compared to using TF-IDF tokens

The experiment results of NNs using KFT for depression detection are shown in Table 7. The RNN, 1DCNN, and BiLSTM increased accuracy, whereas LSTM and GRU still couldn't differentiate the two classes. 1DCNN achieved an accuracy of 78.8%, significantly improved compared to using TF-IDF tokens: but the other NNs still do not perform well using KFT tokens.

Table 7. The evaluation metrics of the NNs using KFT tokens

No.	Model	Accuracy (%)	Precision (%)			Recall (%)			F1 (%)			AUC (%)
			0	1	w. avg.	0	1	w. avg.	0	1	w. avg.	
1	RNN	58.7	58.9	58.2	58.6	76.9	37.4	58.7	66.7	45.6	56.9	61.9
2	1DCNN	**81.6**	**85.3**	**77.9**	**81.9**	79.6	**84.1**	**81.6**	**82.4**	**80.9**	**81.7**	**87.2**
3	LSTM	53.8	53.8	0.0	29.0	**100.0**	0.0	53.8	70.0	0.0	37.6	54.2
4	GRU	53.8	53.8	0.0	29.0	**100.0**	0.0	53.8	70.0	0.0	37.6	50.4
5	BiLSTM	61.2	68.6	56.2	62.4	51.6	72.4	61.2	58.9	63.3	60.9	65.5

Of all the experiments using traditional feature extraction algorithms (Non transfer learning), the best performance was achieved by RF using TF-IFD tokens (See Table 5).

5.2 *Experiments of transfer learning*

In this section, we demonstrate the experiment results of text depression detection using transfer learning. Leveraging the pre-trained BERT-base-uncased, the text contents were first converted into tokens using the pre-trained model's tokenizer-WordPiece tokenization. It breaks down words into smaller units, which helps in handling a large variety of

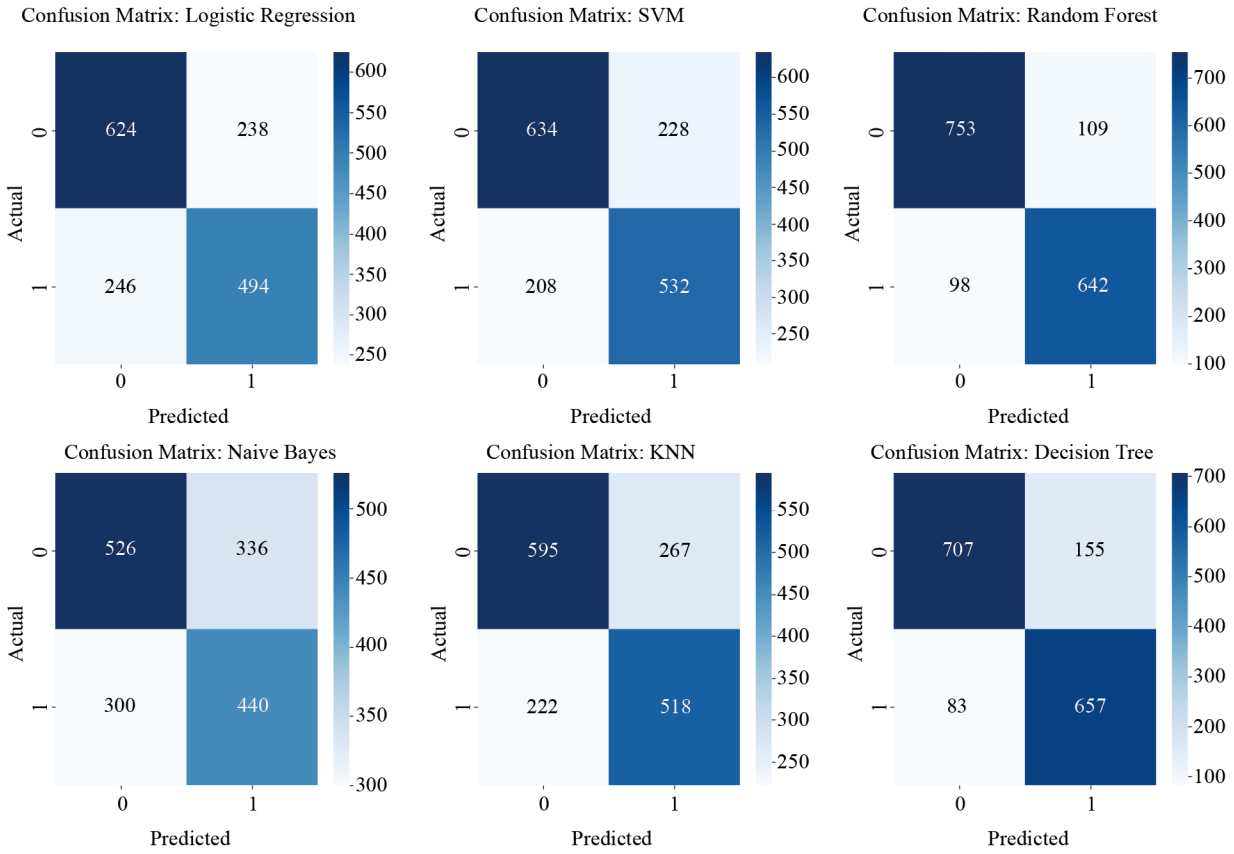
words while maintaining a relatively small vocabulary. For example, the vocabulary doesn't have to include the word "unwilling" by breaking the word "unwilling" down into "un" and "willing", and thus remains a small vocabulary size. After converting to BERT tokens, the pre-trained BERT-base-uncased is called to generate embeddings from these tokens. Then the embeddings are used to train the learning models again to better fit the depression detection task.

For fair comparison, all the previously used conventional ML models and the neural network models were tested for transfer learning too. Table 8 summarises the accuracy, precision, recall, F1 score, and AUC of the conventional MLs using BERT embeddings. The embeddings were used as features extracted by BERT to train these models. The parameter settings for these models remain the same as those in Table 4. In terms of accuracy, all the models increased compared to using KFT tokens but decreased compared to TF-IDF tokens. RF achieved the highest score on all the metrics, except for the precision of class 0 and recall of class 1, the highest of which was achieved by DT. DT scored the second-highest of accuracy. Figure 7 (a) and (b) visually demonstrate the classification results through the confusion matrices of these models and the ROC curves. As can be seen in the confusion matrices, KNN demonstrated a relatively balanced performance compared to TF-IFD tokens; LR, SVM, NB and KNN demonstrated a relatively balanced performance compared to KFT tokens. In the ROC curves, we observe a decrease in AUC on all the ML models compared to TF-IFD tokens. This indicates adopting these conventional ML models in transfer learning does not improve their performance in this task.

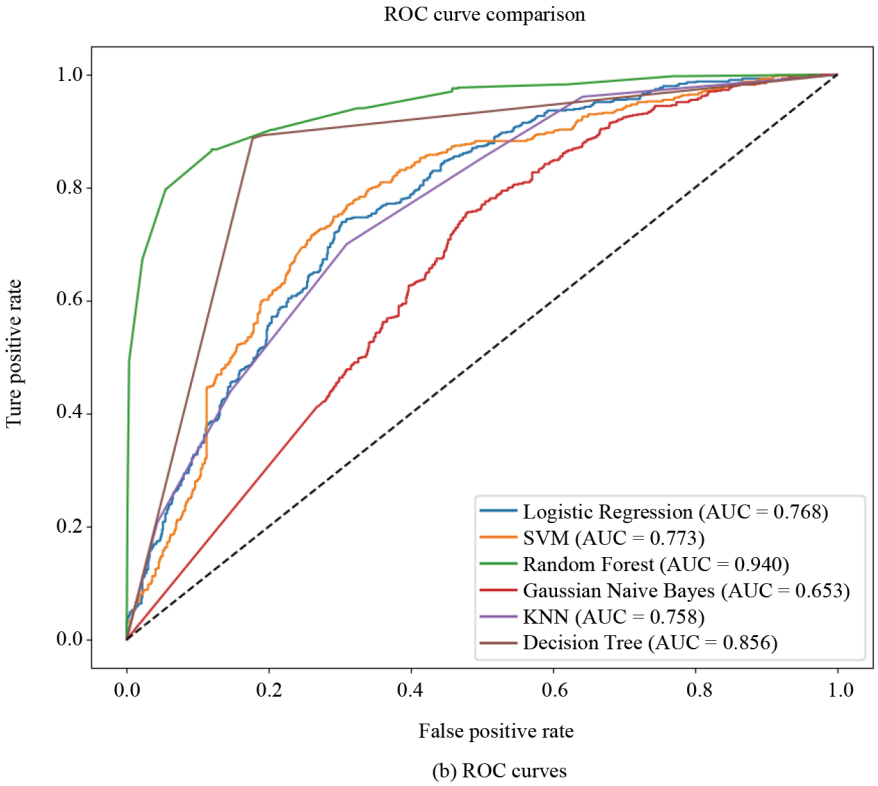
Table 8. The evaluation metrics of the conventional ML models using transfer learning

No.	Model	Accuracy (%)	Precision (%)			Recall (%)			F1 (%)			AUC (%)
			0	1	w. avg.	0	1	w. avg.	0	1	w. avg.	
1	LR	69.8 ↓	71.7	67.5	69.8	72.4	66.8	69.8	72.1 ↓	67.1	69.8	76.8
2	SVM	72.8 ↓	75.3	70.0	72.6	73.5	71.9	72.8	74.4	70.9	72.8	77.3
3	RF	**87.1** ↓	88.5	**85.5**	**87.1**	**87.4**	86.8	**87.1**	**87.9**	**86.1**	**87.1**	**94.0**
4	NB	60.3 ↓	63.7	56.7	60.5	61.0	59.5	60.3	62.3	58.0	60.3	65.3
5	KNN	69.5 ↑	72.8	66.0	69.4	69.0	70.0	69.5	70.9	67.9	69.5	75.8
6	DT	85.1 ↓	**89.5**	80.9	85.5	82.0	**88.8**	85.1	85.6	84.7	85.2	85.6

↓ indicates decrease compared to using TF-IDF tokens, ↑ indicates increase compared to using TF-IDF tokens



(a) Confusion matrices



(b) ROC curves

Figure 7. Visual comparison of the conventional ML model using transfer learning

5.3 *Discussion*

In the experiment of NNs using transfer learning, we observe performance increases on all the five tested NNs. Table 9 summarises the evaluation scores of these experiments. The 1DCNN model achieved the highest scores in terms of accuracy, weighted average precision, weighted average recall, weighted average F1, and AUC. GRU yielded the highest score in the precision of class 0 and recall of class 1.

Table 9. The evaluation metrics of the NNs using transfer learning

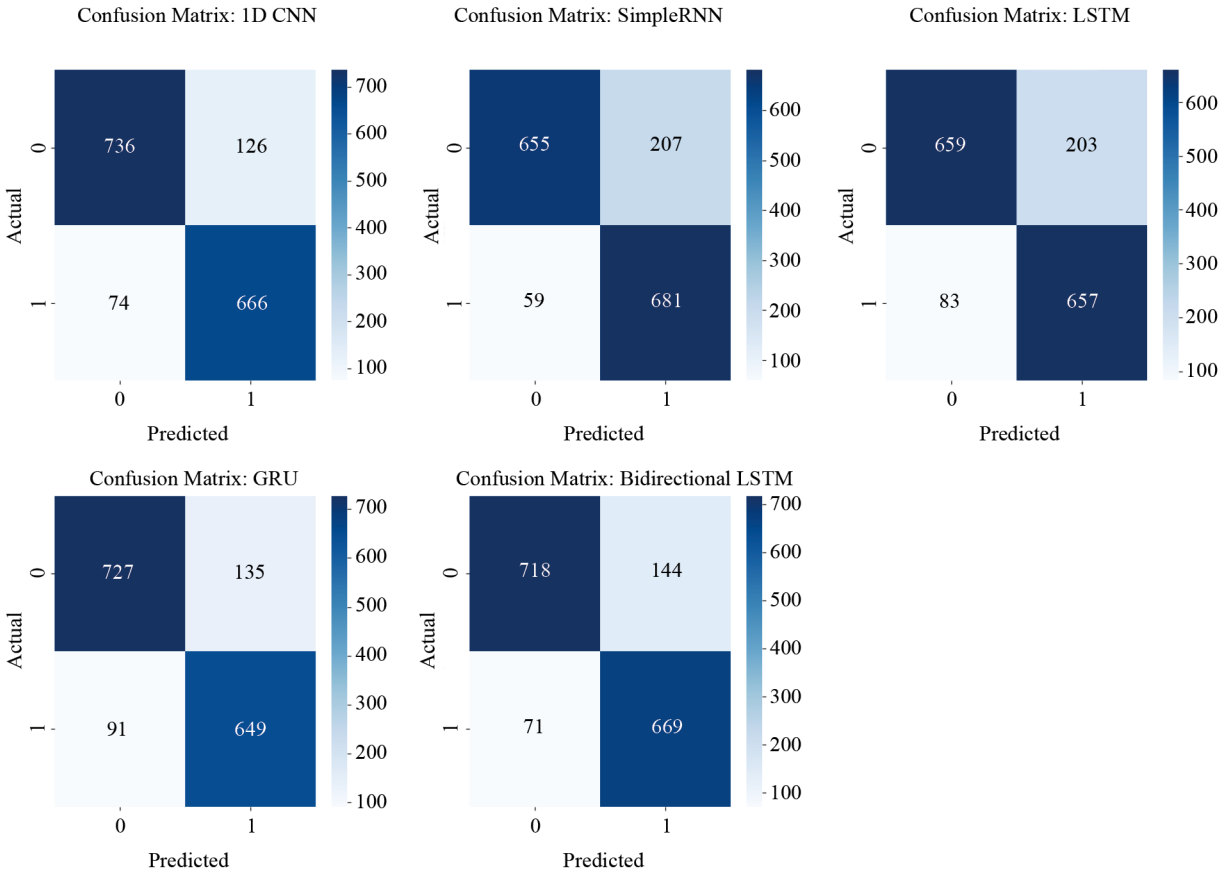
Model	Accuracy (%)	Precision (%)			Recall (%)			F1 (%)			AUC (%)
		0	1	w. avg.	0	1	w. avg.	0	1	w. avg.	
RNN	86.5 ↑	89.8	83.2	86.7	84.6	88.8	86.5	87.1	85.9	86.5	90.4
1DCNN	**89.3**↑	90.0	**88.7**	**89.4**	**90.3**	88.3	**89.3**	**90.1**	**88.4**	**89.4**	**94.8**
LSTM	85.5 ↑	86.2	84.6	85.4	86.9	83.8	85.5	86.5	84.2	85.5	89.5
GRU	85.8 ↑	**90.9**	81.0	86.3	81.8	**90.4**	85.8	86.1	85.4	85.8	89.7
BiLSTM	85.7↑	87.4	83.8	85.7	85.8	85.5	85.7	86.6	84.7	85.7	91.4

The reported scores are the average (arithmetic mean) of 5 tests
↓ indicates decrease compared to using KFT tokens, ↑ indicates increase compared to using KFT tokens

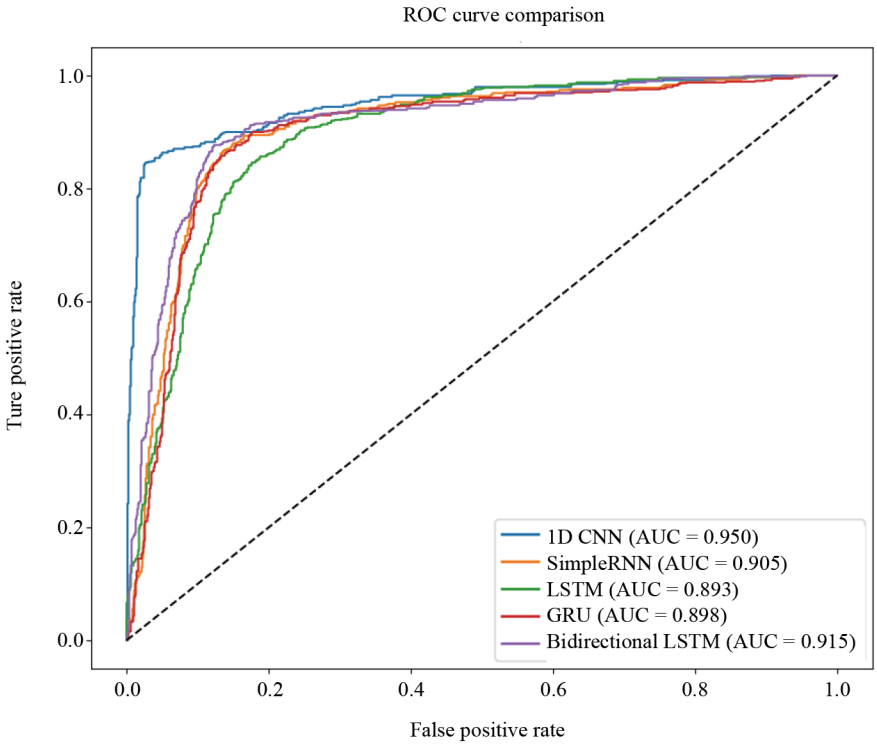
The ROCs are demonstrated in Figure 8. The 1DCNN reached a TPR of 83% at a low FPR near 0.01. The BiLSTM model yielded an F1 of 91.4%, which ranks second.

The train history is demonstrated in Figure 9. The initial training loss of the 1DCNN model was around 0.6 and dropped rapidly within 5 epochs. The initial validation loss was slightly higher than the training loss and did not remarkably decrease as training continued, but the losses were lower than the rest of the models throughout the whole training process. The model with the highest loss was LSTM, which led to the lowest test accuracy of the five NNs (see Table 9).

Figure 10 (a) and (b) demonstrates the effect of using transfer learning on the conventional MLs. The adopting of transfer learning not only improved the KNN's accuracy, but also improved its recall and F1 in terms of balance. But it did not improve the performance of other ML models. As for the NNs, the adoption of transfer learning significantly improved their performance in terms of all the metrics (see Figure 10 (c) and (d)).



(a) Confusion matrices



(b) ROC curves

Figure 8. Visual comparison of the NNs using transfer learning

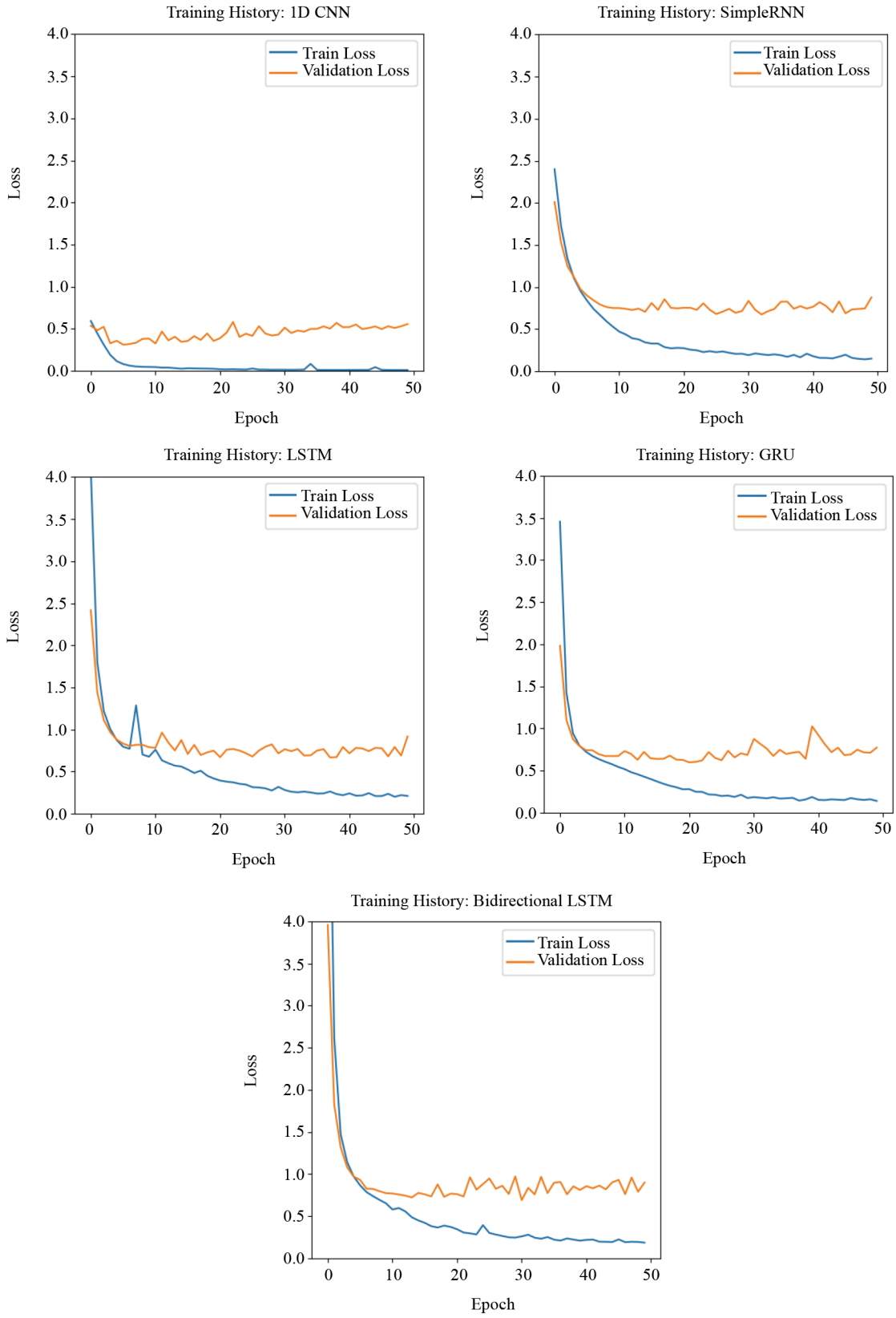
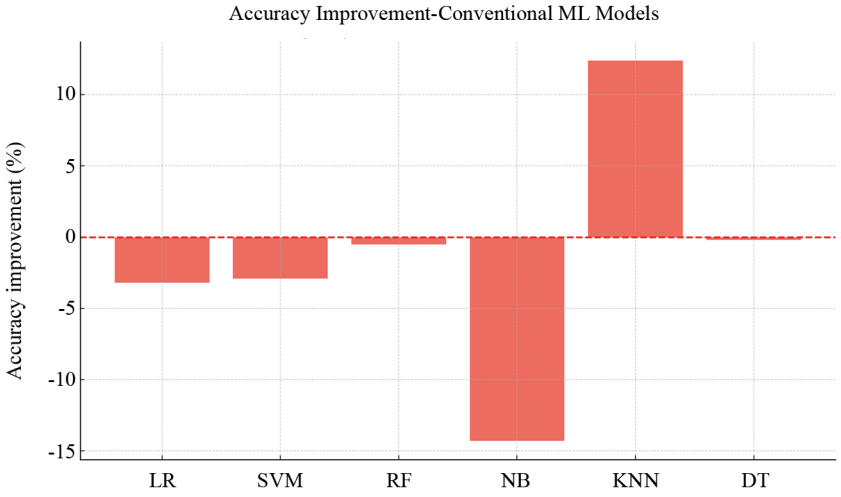
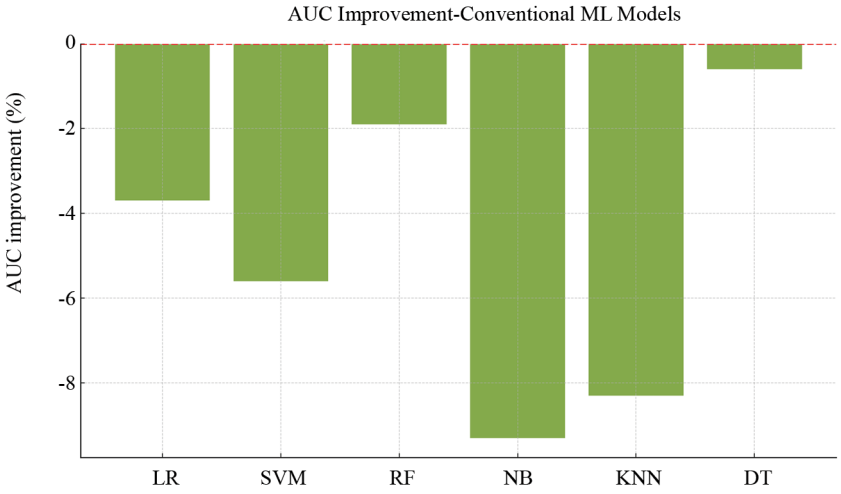


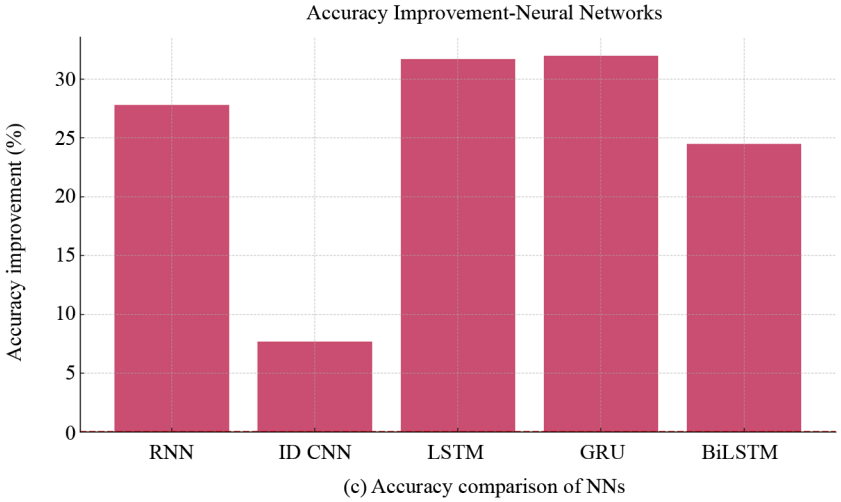
Figure 9. The training history of the NNs using transfer learning



(a) Accuracy comparison of conventional MLs



(b) AUC comparison of convetnional MLs



(c) Accuracy comparison of NNs

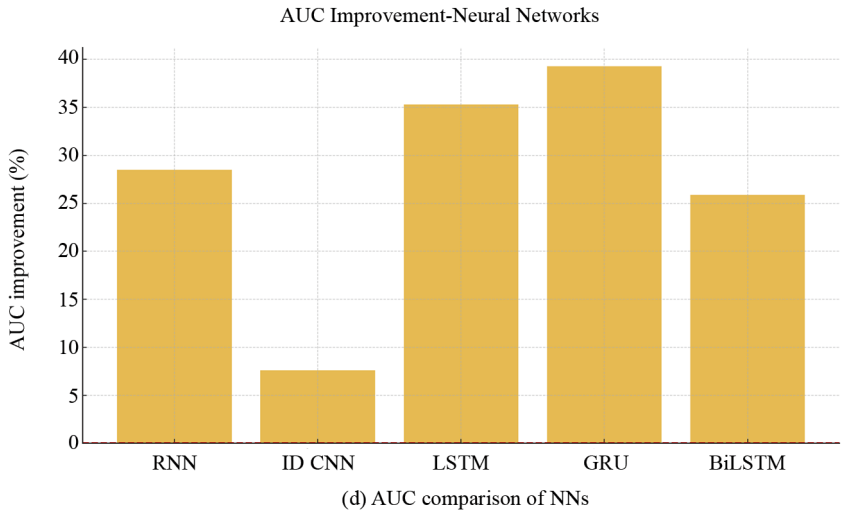


Figure 10. Visual comparison of the adoption of transfer learning

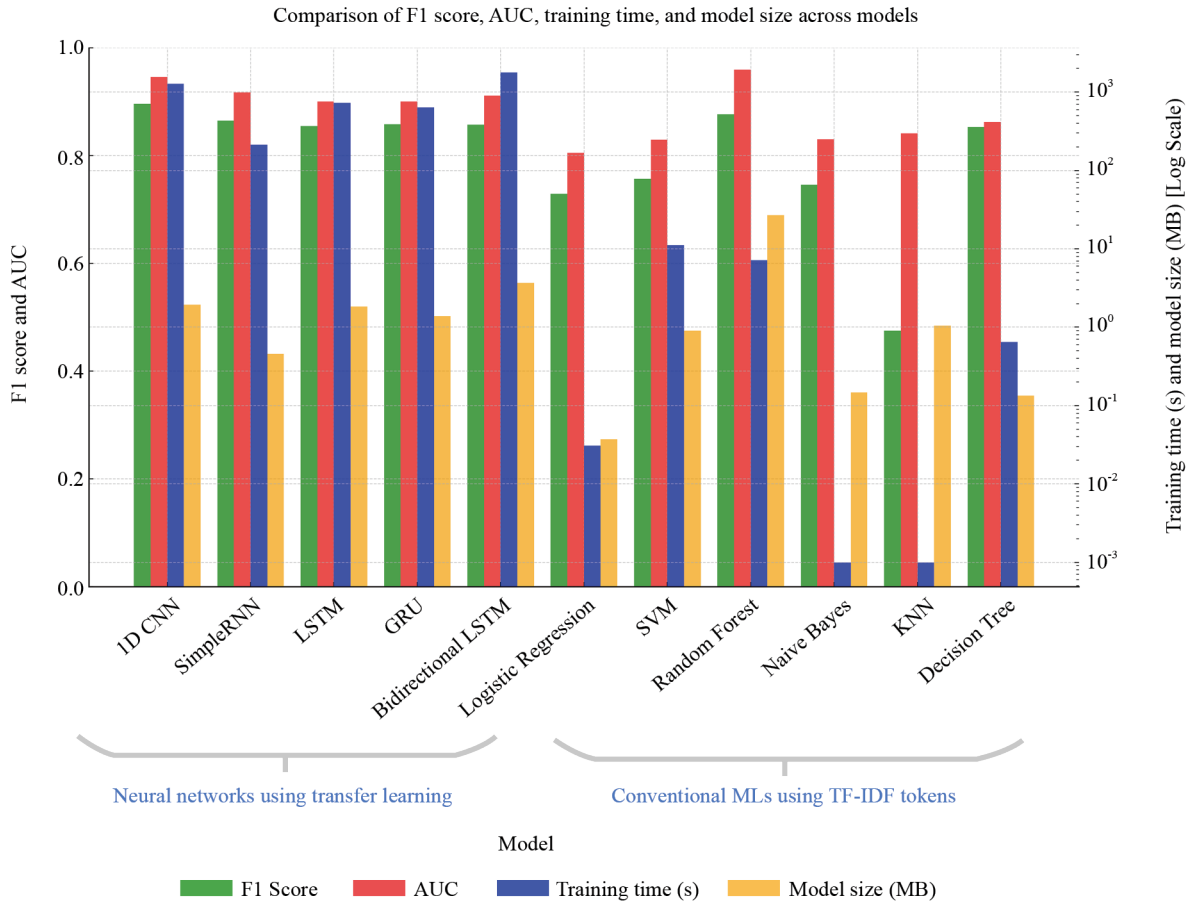


Figure 11. Comparison of F1, AUC, training time and model size across models

In real applications, computation efficiency is one of the key factors engineers must consider. Hence, we select the models that performed well in this task, and compare their training time and model size along with the F1 and AUC

scores, as summarised in Figure 11. Among the tested NNs, 1DCNN and RNN demonstrated higher scores in terms of accuracy (89.3%, 86.5%) and F1 (89.4%, 86.5%). When comparing the accuracy across all models, RF with TF-IDF tokens ranks second (87.6%), higher than RNN using transfer learning, but its model size is bigger than the RNN. Notably, embeddings as extracted features performed better than traditional features, demonstrating their effectiveness in enhancing model performance.

When evaluating the impact of the main parameters on performance, dense units of 128 consistently achieve peak performance, striking an optimal balance between model complexity and performance. For kernel size, values of 3 or 5 generally perform well, effectively capturing both localized and broader contextual patterns, which are crucial for many tasks. Regarding filter size, a size of 128 offers the best trade-off, balancing accuracy and recall while maintaining computational efficiency.

5.4 *Comparison with the state of the art methods*

Text content provides valuable information for depression assessment, therefore, researchers have tried a variety of methods for automatic depression detection using text content.

Table 10. Comparison with the state of the art methods for text depression detection

Study	Model structure	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
[27]	1DCNN, LSTM, Fasttext	Reddit	87	88	87	87
		Twitter	88	89	87	88
[28]	XLNet, CNN, BiGRU	Sina microblog	90.97	**90.87**	87.49	89.15
[29]	Multi-channel 1DCNN Attn.	CLEF-eRisk	**91**	65.4	76.5	70.51
[22]	BERT, T5, Dense	DAIC	89.13	80	85.71	82.76
[30]	BiLSTM	DAIC	-	80	76	78
[20]	S-RoBERTa, BiLSTM	DAIC	-	-	-	80.6
[31]	Multi-layer Perceptron	E-DAIC	-	87	81	83
[32]	BERT, BiLSTM	DAIC	-	83	83	83
[33]	BiLSTM, heterogeneous graph Attn.	DAIC	-	79	80	79
		EATD	-	-	-	-
		CMDC	-	-	-	-
		MODMA	-	-	-	-
Ours	BERT, 1DCNN	E-DAIC	89.3	89.4	89.3	89.4
		Mental Health Corpus	89.5	90.7	88.2	89.4

Attn.: to Attention mechanism

Table 10 summarises the state-of-the-art methods of text depression detection in recent years. As can be seen, some studies were conducted on datasets collected through social media platforms such as Reddit, Twitter, Sina Microblog, etc., because collecting clinically diagnosed depression data is difficult due to many reasons such as privacy concerns and limited medical resources. Among the recent studies, CNN (usually 1DCNN for text data) and BiLSTM are the most frequently adopted models. This is consistent with the results of this study, as the highest and second highest F1 scores were achieved by the 1DCNN and BiLSTM models. GRU and LSTM are also adopted in some studies as they are common for sequential data.

6. Conclusion

In the field of health care and medicine, high-quality annotated data require professional expertise in diagnosis. To ensure an AI model accurately captures the complex and diverse characteristics of patients, the training data should be

collected from clinically diagnosed patients, which further increases the cost and difficulties of data collection. Hence, in the field of AI-driven depression detection, researchers face the difficulties of insufficient training data. This article reports the study of enhancing depression detection of AI models trained on a small dataset by adopting transfer learning technology. 11 models, including 6 conventional machine learning models and 5 neural networks were tested under comparison. Based on the experiments results, the following conclusions are drawn.

• With the TF-IDF and KTF tokens, the conventional MLs performed well, whereas NNs did not demonstrate satisfying results.

• The RF reaches a high TPR at an early stage (FPR < 0.01), and increases slowly after that, leading to a higher AUC.

• Some models demonstrated poor balance. For example KNN (recall of 95.8% on class 0, 12.0% on class 1) and all the NNs with TF-IDF, LR, SVM and NB with KFT.

• Transfer learning improved the performance in this task for the neural networks, but did not for the conventional MLs.

• The best performance of all the tested models was achieved by the 1DCNN model trained using transfer learning technique, the adoption of transfer learning increased the accuracy by 7.7%.

Future work could focus on incorporating interpretability techniques to make deep learning models more transparent and clinically relevant. By doing so, the model could offer insights into its decision-making process, thereby building trust and usability in sensitive clinical domains. Additionally, research could explore the development of more complex models that are adaptable across different languages, addressing the challenges of linguistic and cultural variability in depression detection. Future work could also include detecting multiple levels of depression severity to enhance the model's applicability in diverse clinical scenarios.

Data availability

The data used in this study is available for request at https://dcapswoz.ict.usc.edu/.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Semrau M, Alem A, Ayuso-Mateos JL, Chisholm D, Gureje O, Hanlon C, et al. Strengthening mental health systems in low-and middle-income countries: recommendations from the Emerald programme. *BJPsych Open*. 2019; 5(5): e73.

[2] Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck depression inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment*. 1996; 67(3): 588-597. Available from: https://doi.org/10.1207/s15327752jpa6703_13.

[3] Zung WW. A self-rating depression scale. *Archives of General Psychiatry*. 1965; 12(1): 63-70.

[4] Hamilton M. The hamilton rating scale for depression. In: *Assessment of Depression*. Berlin, Heidelberg: Springer; 1986. p.143-152. Available from: https://doi.org/10.1007/978-3-642-70486-4_14.

[5] Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of Affective Disorders*. 2001; 66(2-3): 159-164.

[6] Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA*. 1999; 282(18): 1737-1744.

[7] Adam-Troian J, Bonetto E, Arciszewski T. Using absolutist word frequency from online searches to measure population mental health dynamics. *Scientific Reports*. 2022; 12: 2619.

[8] Liu T, Meyerhoff J, Eichstaedt JC, Karr CJ, Kaiser SM, Kording KP, et al. The relationship between text message sentiment and self-reported depression. *Journal of Affective Disorders*. 2022; 302: 7-14.

[9] Kim Y. Convolutional neural networks for sentence classification. In: Moschitti A, Pang B, Daelemans W. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics; 2014. p.1746-1751.

[10] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997; 9(8): 1735-1780.

[11] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*. 2005; 18(5-6): 602-610.

[12] Ahmad Wani M, ELAffendi MA, Shakil KA, Shariq Imran A, Abd El-Latif AA. Depression screening in humans with AI and deep learning techniques. *IEEE Transactions on Computational Social Systems*. 2023; 10(4): 2074-2089.

[13] Church KW. Word2Vec. *Natural Language Engineering*. 2017; 23(1): 155-162.

[14] Baena-García M, Carmona-Cejudo JM, Castillo G, Morales-Bueno R. TF-SIDF: Term frequency, sketched inverse document frequency. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE; 2011. p.1044-1049.

[15] Kratzwald B, Ilić S, Kraus M, Feuerriegel S, Prendinger H. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*. 2018; 115: 24-35. Available from: https://doi.org/10.1016/j.dss.2018.09.002.

[16] Choudhry A, Khatri I, Jain M, Vishwakarma DK. An emotion-aware multitask approach to fake news and rumor detection using transfer learning. *IEEE Transactions on Computational Social Systems*. 2024; 11(1): 588-599.

[17] Ghosh S, Priyankar A, Ekbal A, Bhattacharyya P. Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. *Knowledge-Based Systems*. 2023; 260: 110182. Available from: https://doi.org/10.1016/j.knosys.2022.110182.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017; 30: 1-11.

[19] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE; 2015. p.19-27.

[20] Milintsevich K, Sirts K, Dias G. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*. 2023; 10(1): 4.

[21] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:190711692*. 2019. Available from: https://doi.org/10.48550/arXiv.1907.11692.

[22] Zhang J, Guo Y. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognition Letters*. 2024; 178: 167-173. Available from: https://doi.org/10.1016/j.patrec.2024.01.005.

[23] Hadikhah Mozhdehi M, Eftekhari Moghadam A. Textual emotion detection utilizing a transfer learning approach. *The Journal of Supercomputing*. 2023; 79(12): 13075-13089.

[24] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:181004805*. 2018. Available from: https://doi.org/10.48550/arXiv.1810.04805.

[25] Bounabi M, El Moutaouakil K, Satori K. Text classification using fuzzy TF-IDF and machine learning models. In: *Proceedings of the 4th International Conference on Big Data and Internet of Things*. New York, United States: Association for Computing Machinery; 2019. p.1-6.

[26] Shorten C, Khoshgoftaar TM. Language models for deep learning programming: A case study with Keras. In: *Deep Learning Applications, Volume 4*. Springer; 2022. p.135-161.

[27] Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2024; 23(1): 1-20.

[28] Wang L, Zhang Y, Zhou B, Cao S, Hu K, Tan Y. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. *Computers and Electrical Engineering*. 2024; 118: 109413. Available from: https://doi.org/10.1016/j.compeleceng.2024.109413.

[29] Dalal S, Jain S, Dave M. Convolution neural network having multiple channels with own attention layer for depression detection from social data. *New Generation Computing*. 2024; 42(1): 135-155.
[30] Iyortsuun NK, Kim SH, Yang HJ, Kim SW, Jhon M. Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. *IEEE Access*. 2024; 12: 20479-20489.
[31] Villatoro-Tello E, Ramírez-de-la Rosa G, Gática-Pérez D, Magimai Doss M, Jiménez-Salazar H. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In: *Proceedings of the 2021 International Conference on Multimodal Interaction. ICMI '21*. New York, NY, USA: Association for Computing Machinery; 2021. p.557-566.
[32] Rai BK, Jain I, Tiwari B, Saxena A. Multimodal mental state analysis. *Health Services and Outcomes Research Methodology*. 2024. Available from: https://doi.org/10.1007/s10742-024-00329-2.
[33] Chen T, Guo Y, Hao S, Hong R. Semi-supervised domain adaptation for major depressive disorder detection. *IEEE Transactions on Multimedia*. 2024; 26: 3567-3579.