Detection of Depression Among Arabic Twitter Users Using a Convolutional Neural Network

Mohammed Majid Msallam[1,*], Esraa Yahia Tarkan[2], Sarah Kareem Salim[3]

[1]College of Artificial Intelligence Engineering, University of Technology, Baghdad, Iraq

[2]Ministry of Higher Education and Scientific Research, Baghdad, Iraq

[3]Electrical Engineering Department, University of Misan, Misan, Iraq

Received 26 December 2025; received in revised form 11 July 2026; accepted 13 July 2026

DOI: https://doi.org/10.46604/peti.2026.16024

Abstract

Depression is a prevalent mental health condition, where early detection using the analysis of patients' feelings contributes to preventing and treating the exacerbation of symptoms. This work aims to detect depression in Arabic tweets using convolutional neural networks (CNN) trained on a dataset generated from the Arabic_Dep_tweets_10,000 dataset. Two datasets are generated using the stripping and captioning process to improve the generalization of the proposed CNN model. The stripping process removes punctuation marks from Arabic tweets, while the captioning process types the Arabic tweet on a white background to generate an image. Experimental results demonstrate that the proposed CNN model achieves up to 100% accuracy on the generated dataset treated with stripping and the captioning process.

Keywords: convolutional neural network, depression, prevention, Arabic tweets, mental health

1. Introduction

Millions of people worldwide suffer from mental health conditions, including anxiety and depression, making mental health an essential component of total well-being. The World Health Organization (WHO) defines depression as a common mental health condition marked by feelings of guilt or low self-worth, loss of interest or pleasure, fatigue, poor attention, and disturbed sleep or eating [1]. Detecting depression in its early stages is important, as more than 75% of people remain undiagnosed, and their illnesses worsen [2]. Individuals who do not acquire therapy for depression promptly face worsening symptoms. Therefore, depression is considered a dangerous disease that not only affects the mental state of a person but also causes physical harm to a patient. The severity of depression is predicted in terms of the mental health condition of a patient. According to a survey conducted by the WHO, approximately 280 million people suffer from depression worldwide, and nearly 800000 cases of suicide due to depression are reported every year.

Researchers estimate that nearly one in five people experience depression at least once in their lifetime. With the ever-increasing number of people suffering from depression, the need for early detection of this mental health condition has become increasingly important to mitigate its negative effects and improve treatment outcomes. Therefore, the need arises to develop automated systems capable of detecting early signs of depression and providing the right support to those affected [3]. Traditional methods for evaluating depression, including clinical interviews, psychological rating scales, and self-reported questionnaires, are widely used to identify depressive symptoms and measure their severity. The most familiar of these tools are the Hamilton depression rating scale, the beck depression inventory (BDI), and the patient health questionnaire 8 (PHQ-8) score, in addition to structured clinical interviews designed to evaluate the severity of depressive symptoms and behaviors.

* Corresponding author. E-mail address: mohammed.m.arjeeli@uotechnology.edu.iq

These methods often require direct patient participation, professional evaluation, and ongoing follow-up, which make them time-consuming and costly. The social shame associated with mental health disorders may lead some individuals to conceal their true mental state or hesitate to show it, which can result in inaccurate or incomplete assessments in clinical interviews [4].

According to these challenges, scientists have concentrated on exploring alternative sources that can help in the early detection of depression more efficiently. Social media has appeared to be one of the most important of these sources, given its increasing use in daily life. Thoughts and feelings of people are declared from social media platforms according to their activities on these media. Thoughts in posts are analyzed to help us to detect and diagnose some psychological diseases such as depression [5]. Some recent research utilizes posts and tweets on social media as a valuable resource to detect symptoms and signs of depression disorders. New approaches exist for detecting depression using artificial intelligence (AI) mechanisms and advanced learning methods [6].

As data generation grows rapidly through these platforms, the use of AI and deep learning (DL) techniques to analyze these data and extract patterns associated with depression has increased. One of the most advanced and effective branches of AI is the DL models, especially convolutional neural network (CNN), which enable automatic learning to create hierarchical representations of features from large and complex datasets; this model consists of many artificial neurons. DL can analyze data and extract useful information from complex patterns in a dataset [7]. Many DL models have been developed across various fields, including image and disease classification [8]. Thereby, DL is used to detect depression by analyzing and extracting features from tweets using a CNN model [9].

Many studies have used DL models to detect depression in social media tweets in various languages, including English, Spanish, French, and Arabic [10-11]. Detecting depression in Arabic tweets is still a significant challenge. Unlike English, Arabic is characterized by its rich morphology, diverse dialects, and limited linguistic resources, which can negatively affect the efficiency of natural language processing (NLP) techniques and traditional feature extraction methods [12-13].

Most current methods rely heavily on text preprocessing and linguistic characteristics extraction. However, the performance of these methods can be affected by the complexity of the Arabic language and the variety of writing styles used in Arabic texts. To reduce reliance on language-specific feature engineering, this study explores an alternative approach to Arabic tweets by converting them into images and using a CNN model to detect depression. By learning visual patterns directly from tweet representations, the proposed approach aims to provide a more robust and effective solution for detecting depression in Arabic Twitter data. Therefore, the goal of this study is to develop a CNN model for detecting depression in Arabic tweets accurately. The key contributions of this research are summarized as follows:

(1) Proposing a model based on CNNs for detecting depression in Arabic tweets.

(2) Presenting an alternative method for representing Arabic tweets by converting them into images, thus reducing reliance on traditional linguistic feature extraction techniques used in NLP.

(3) Enhancing the generalizability of the proposed model by removing ineffective characters from Arabic tweets during the preprocessing stage.

(4) Evaluating the performance of the proposed model on test data using the accuracy and F1-score metrics.

The rest of this work is prepared as follows. Section 2 presents related articles for the proposed work, while Section 3 shows the methodology consisting of a dataset and a proposed model based on CNN. Then, Section 4 represents the experimental results, and Section 5 concludes this article.

2. Literature Survey

Various related works discuss the detection of depression using different approaches. Lyu et al. built a text-only prediction model for detecting depression in Chinese microbiology users by expanding on traditional psycholinguistic features to include cultural psychological factors and suicide-related lexicons [14]. Their method involved collecting self-reported center for epidemiological studies depression scale (CES-D) scores and past Weibo posts from 789 users. They extracted 117 lexical features using the simplified Chinese linguistic inquiry and word count (SCLIWC) alongside Chinese versions of the suicide, moral motivation, moral foundations, and individualism and collectivism dictionaries. Villa-Pérez et al. collected two new datasets adapted for Spanish and English in the field of automatic mental state recognition [15]. The datasets include 1,500 users who self-reported a confirmed diagnosis of mental illness and 1,700 undiagnosed users. The researchers used multiple types of text features such as n-grams, q-grams, part-of-speech (POS) tags, topic modeling, linguistic inquiry and word count (LIWC), and word embeddings to train and compare traditional machine learning (ML) and DL models. They found that XGBoost and CNN, using n-grams of words from one to three, deliver the optimal performance, with lexical features demonstrating outstanding effectiveness.

Cha et al. analyzed three separate datasets to identify terms that aid in diagnosing depression in the Korean, Japanese, and English languages [16]. Their work presented a lexicon-based, DL approach for detecting depression in social media posts, addressing the need for early detection and the gap in multilingual resources. Their models were built on CNN, bidirectional long short-term memory (BiLSTM), and bidirectional encoder representations from transformers (BERT). Each language has characteristics that are influenced by the culture of its speakers, where these characteristics affect model performance [17].

Most models in the tweet's depression detection are developed by researchers based on multiple English datasets. The study by Khan and Alqahtani, which aimed to detect depressive signs through sentiment analysis of tweets. BERT, term frequency inverse document frequency (TF-IDF), and spacy were used to extract feature, tested and evaluated four hybrid ML models. The results indicated that the combination of TF-IDF and logistic regression achieved optimal performance, making it a reliable tool suited for early depression detection based on tweets [18]. Zhou and Mohd addressed the challenge of detecting depression in tweet texts, noting that traditional techniques often do not interpret informal and ambiguous language, including emojis and slang, common on social platforms [19]. Their framework integrated advanced text preprocessing with a DL model to overcome this limitation. Their model proves high performance compared by traditional baselines in detecting depression in English tweets dataset. However, these models perform poorly when applied to the Arabic dataset due to differences in context between the Arabic and English languages.

In detecting depression in Arabic tweets, researchers designed the models using DL and ML. Abdulsalam et al. addressed the critical gap in automated suicide detection within Arabic-language social media content using ML and DL techniques [20]. The primary contribution of their work included the creation and expert annotation of the first Arabic suicidal tweet dataset, including 5,719 tweets collected from Twitter. Their study investigated pre-trained DL models, proving that the AraBert transformer model significantly outperforms all other methods, achieving an 88% F1-score and 91% accuracy. In another work, Rabie et al. presented a novel framework for detecting major depressive disorder by analyzing Arabic user-generated content from Twitter [21]. They compiled and preprocessed a dataset from three sources to address the high prevalence of depression and the relative scarcity of studies on Arabic text. Their study evaluated ML models including support vector machine (SVM), random forest (RF), logistic regression (LR), and Gaussian naive Bayes (GNB), and DL models including a LSTM model and BERT transformers. Their findings provided a highly accurate and rapid method for detecting depression from tweets, potentially aiding in early detection. These models lack high accuracy to detect depression in Arabic tweets. Therefore, a need exists for a model that performs well to detect early depression to help with intervention and prevent the worsening of its symptoms.

Table 1 summarizes the related works where the researched gap appears clearly. The accuracy of the CNN model in detecting depression in Arabic tweets needs to improve. Thus, the key goals of the proposed approach are to build a model that achieves high accuracy in detecting depression in Arabic tweets. This goal is realized by extracting efficient and helpful features to ensure high accuracy for the trained model.

Table 1 Summary of related works

Ref.	Year	Best model	Language	Metrics	Score
S. Lyu et al[14]	2023	Linear Regression	Chinese	R-squared	0.1000
M. E. Villa-Pérez et al[15]	2023	XGBoost and CNN	Spanish	Accuracy	0.8350
M. E. Villa-Pérez et al[15]	2023	XGBoost and CNN	English	Accuracy	0.8460
J. Cha et al[16]	2022	BiLSTM	Korean	Accuracy	0.9991
J. Cha et al[16]	2022	BERT	Japanese	Accuracy	0.9993
J. Cha et al[16]	2022	BERT	English	Accuracy	0.9991
K. M. Broczek et al[17]	2024	Logistic Regression	English	Accuracy	0.9940
S. Zhou and M. Mohd[19]	2025	BERT-BiLSTM	English	Accuracy	0.8895
A. Abdulsalam et al[20]	2024	AraBert	Arabic	Accuracy	0.9100
E. M. Rabie et al[21]	2025	BERT transformers	Arabic	Accuracy	0.9303

This research aims to address the critical gap in detecting depression in Arabic tweets by proposing a novel image-based textual representation framework. Specifically, the proposed method uses CNN architecture to construct a robust classifier, which significantly outperforms established baseline models. By converting textual data into visual representations, this work introduces an innovative paradigm to the field of mental health detection, circumventing traditional NLP bottlenecks, which stem from complex morphology, dialectal variations, and spelling noise, and yielding highly discriminative spatial features that optimize both the training and evaluation phases of the predictive model.

3. Methodology

Detecting depression in Arabic tweets can be achieved using a variety of NLP with ML techniques. The proposed model for detecting depression in Arabic tweets is based on a CNN, which is a type of DL. DL uses artificial neural networks to uncover intricate links and patterns in data. CNN can process large amounts of data and recognize intricate feature correlations. The proposed approach consists of four key stages, which are preprocessing, splitting, modeling, and evaluation, as shown in Fig. 1. Preprocessing is the first and crucial stage where a new dataset of images will be created from the text in Arabic tweets. Splitting is the process of dividing the preprocessed dataset into three distinct sets, which are the training set, validation set, and testing set. Modeling focuses on the creation and use of the CNN Model. Evaluation is the final step where the performance of the Model is rigorously assessed.

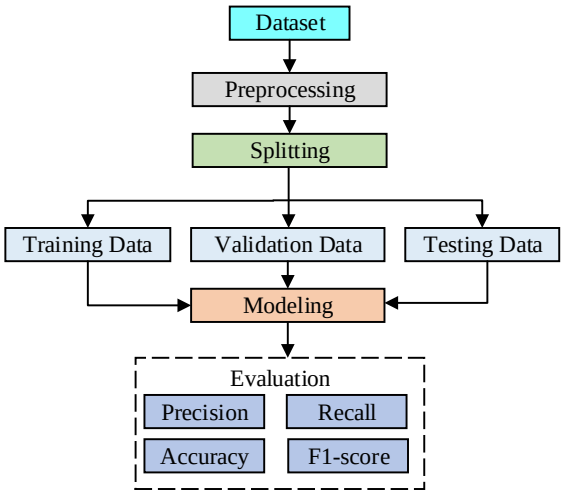


Fig. 1 Overview of the proposed experimental pipeline

3.1. Dataset

This work is based on the Arabic_Dep_tweets_10,000 dataset, which Khalil et al. established [22]. The tweets in this dataset were obtained through the Twitter application programming interface (API). Arabic Tweets in the dataset were collected between January 2019 and April 2022. Arabic tweets in the dataset consist of a mix of Egyptian dialect, Gulf dialects, and modern standard Arabic (MSA). A total of 10,000 Arabic tweets are collected, which contain 5000 tweets labeled as depressed and 5000 tweets annotated as non-depressed, as shown in Fig. 2.

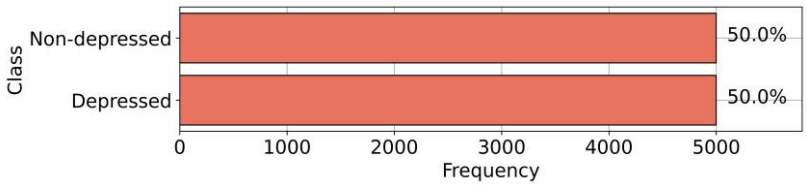
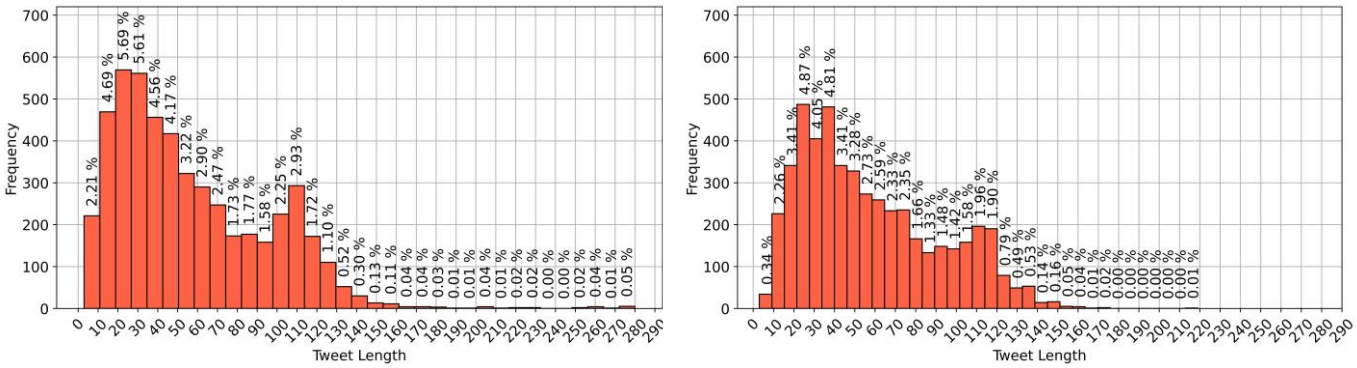


Fig. 2 The classes in the dataset

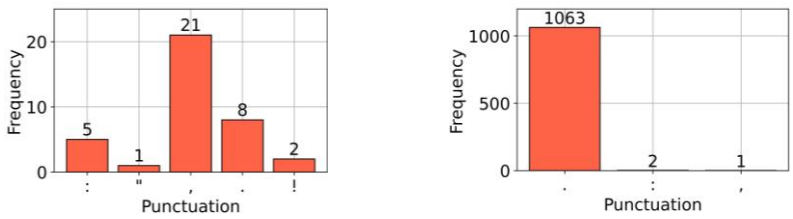
The length of the dataset in the tweets labelled as depressed ranges between 3 and 280 characters. However, the average length of the dataset in the tweets labelled as depressed is approximately 57 characters. The proportion of depressed tweets with a length between 20 and 30 characters accounts for 5.6% of the dataset, as shown in Fig. 3(a). Conversely, the length of the tweets categorized as non-depressed ranges between 3 and 219 characters. However, the average length of the non-depressed tweets is approximately 58 characters, with 4% of this subset falling between 20 and 40 character-lengths, as shown in Fig. 3(b). Both the depression and non-depression classes contain approximately the same count of characters. This feature enables the proposed model to classify correctly.



(a) Tweets labelled as depressed (b) Tweets labelled as non-depressed

Fig. 3 The length of tweets

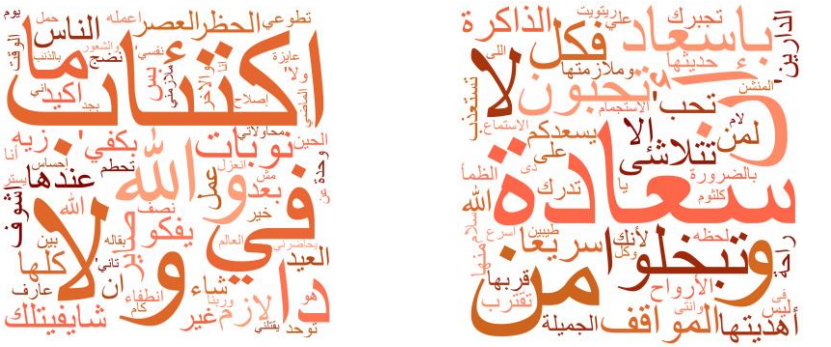
Most tweets contain punctuation, which includes a colon, quotation marks, commas, a period, and an exclamation mark. Fig. 4 shows the punctuation used in the tweets labelled as depressed and tweets labelled as non-depressed in the dataset. Specifically, two tweets labelled as depressed include exclamation mark, whereas no tweet labelled as non-depressed include exclamation mark. There are 1063 periods that appear in non-depressed tweets, compared to only 8 periods used in depressed tweets. However, the frequency of punctuation marks is higher in depressed tweets than in normal behavior tweets. This variation in punctuation usage affects the generalization process of the proposed model. For this reason, the model is trained on two types of data, one containing punctuation and the other devoid of punctuation.



(a) Tweets labelled as depressed (b) Tweets labelled as non-depressed

Fig. 4 Punctuation in the dataset

Fig. 5 demonstrates the word cloud for non-depressed and depressed classes. Less frequently used Arabic words appear in a smaller font, while the most used Arabic terms are displayed in a bigger font. This shows that people with depression use specific vocabulary to express their feelings, a pattern similarly observed among non-depressed people.



(a) Tweets labelled as depressed (b) Tweets labelled as non-depressed

Fig. 5 Word clouds in the dataset

3.2. Preprocessing

The preprocessing stage generates two derivative datasets, dataset I and dataset II, as shown in Fig. 6. In dataset I, the text of the tweet is processed using the captioning process only. In contrast, dataset II is processed using both stripping and captioning process. These datasets help to improve the generalization of the proposed model on unseen data.

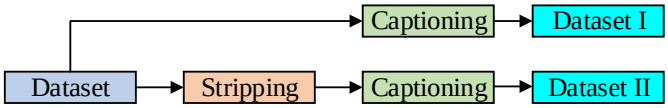


Fig. 6 Preprocess for the dataset

In the captioning process, the Arabic tweet is typed on a white background to generate an image. The Arabic tweets are rendered using the python imaging library (PIL/Pillow) with unicode transformation format (UTF)-8-character encoding to properly support Arabic script. Specifically, the text is typed onto a solid white background using the arial true type font (arial.ttf) at a fixed font size of 20 points, which ensures consistent character spacing, line height, and geometric representation across the entire generated image corpus.

The dimensions of the generated images are 25 × 2000 pixels, and the images are colored. The choice of these dimensions is directly derived from empirical text in Arabic tweet statistics and hardware optimization constraints. Specifically, a width of 2000 pixels is determined based on the maximum character count observed in the Arabic tweet corpus to prevent any data truncation, ensuring that even the longest tweets are fully represented within a single image. A height of 25 pixels is selected to accommodate the character embedding matrix height and formatting padding, while the 3 channels correspond to the standard red, green, and blue (RGB) color space to leverage pre-trained computer vision architectures. The size of the generated image is between 1 kilobyte and 16 kilobytes in Joint Photographic Experts Group (JPEG) format. In the stripping process, punctuation marks are removed from Arabic tweets to enhance the generalization process of the proposed model.

3.3. CNN

A CNN is specifically a kind of deep learning algorithm designed for data with grid patterns, such as images [23]. CNN was created to adaptively and automatically learn spatial feature hierarchies and was inspired by the structure of the animal visual cortex [24]. Typically, a CNN consists of three different kinds of layers, often known as building blocks, which are a convolution layer, a pooling layer, and a fully connected layer [25]. The first two layers, the convolution layer and the pooling layer, function as feature extraction, while the third layer, a fully connected layer, relates the extracted features to predict the output.

A convolution layer is an essential component of a CNN, which is composed of several mathematical operations, including convolution, a particular kind of linear operation. CNNs are very effective for processing images because a feature can appear anywhere in the image. CNNs process digital images using a small grid of parameters called a kernel, which is an optimizable feature extractor. The kernel is applied to each pixel value in a digital image, which is stored as a two-dimensional (2D) grid. The complexity of the retrieved characteristics increases gradually and hierarchically when the output of a layer feeds into the next layer [26]. The process of fine-tuning parameters, the kernels, is performed in training using optimization algorithms such as gradient descent, backpropagation, or other kinds of optimization algorithms. The training reduces the discrepancy between predicted and truth labels.

A CNN model designed to detect Arabic depressed users from Twitter features a sequential arrangement of layers, as shown in Fig. 7. The input Layer serves as the initial architectural component, tasked with ingesting the raw image tensor into the proposed CNN model. The following layers used in the proposed CNN model are a convolutional 2-dimensional (Conv2D), max pooling 2-dimensional (MaxPooling2D), dropout, flatten, and dense layers, where each layer serves a distinct function within the proposed CNN model. The foundational building block of the proposed CNN model is the Conv2D layer, which automatically extracts spatial hierarchies of features by convolving learnable filters from the raw image tensor to create a set of feature maps.

The function of the MaxPooling2D layer is to downsample the feature maps, where it chooses the maximum value within a defined region. The integration of the MaxPooling2D layer into the proposed CNN model is justified by its capacity to reduce the parameter count and computational load significantly. Furthermore, this layer introduces a degree of representation invariance, thereby enhancing the network's robustness against minor translational shifts in the input feature maps. The dropout layer is a regularization technique that randomly sets a fraction of input units to zero during training to prevent overfitting by reducing the interdependence of neurons. The function of the flatten layer is to bridge the gap between the feature extraction portion of the preceding layers and the final classification part by resizing the multi-dimensional output from the dropout layer into a one-dimensional vector. This flattened vector is then sent to the dense layer, which uses it to complete the last classification operation.

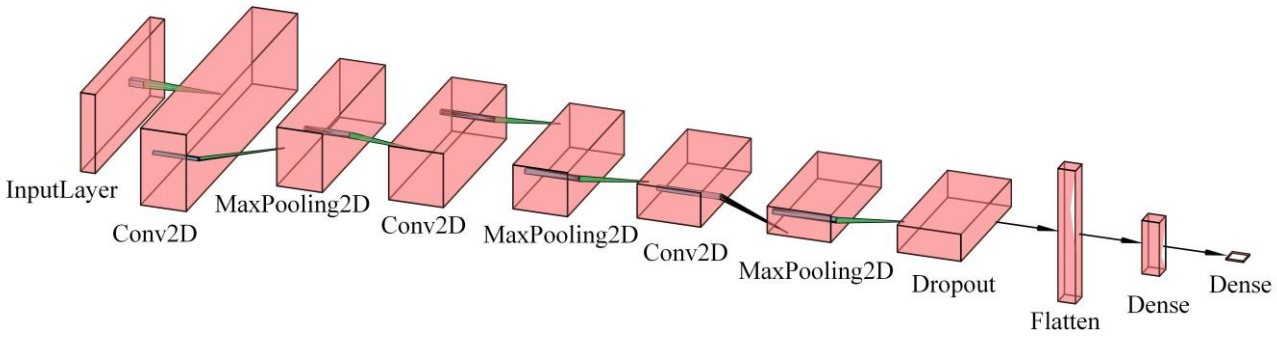


Fig. 7 The proposed CNN model architecture

The dimensions of the generated image are 25 × 2000 × 3. Thereby, the first layer of the proposed model is an input layer with dimensions of 25 × 2000 × 3. The dimensions of hidden layers, from layer_1 to layer_7, are illustrated in Table 2. The activation function of Conv2D and dense layers in the hidden layers is rectified linear unit (ReLU).

The ReLU gives outputs as the input directly if it is positive, and zero otherwise, as mathematically defined [27]:

$$\mathrm{ReLU}(x) = \max(0,\ x) \tag{1}$$

where x denotes the input value to the ReLU function, ReLU (x).

The output layer gives the probability of depression tweet using the sigmoid activation function. The sigmoid function produces an S-shaped curve, as mathematically defined [27]:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

where x denotes the input value to the sigmoid function.

Table 2 The details of the proposed model layers

Name layer	Kind layer	Activation function	Input size	Output size
input	InputLayer	---	25 × 2000 × 3	25 × 2000 × 3
layer_1	Conv2D	ReLU	25 × 2000 × 3	23 × 1998 × 32
layer_2	MaxPooling2D	---	23 × 1998 × 32	11 × 999 × 32
layer_3	Conv2D	ReLU	11 × 999 × 32	9 × 997 × 64
layer_4	MaxPooling2D	---	9 × 997 × 64	4 × 498 × 64
layer_5	Conv2D	ReLU	4 × 498 × 64	2 × 496 × 128
layer_6	MaxPooling2D	---	2 × 496 × 128	1 × 248 × 128
layer_7	Dropout	---	1 × 248 × 128	1 × 248 × 128
layer_8	Flatten	---	1 × 248 × 128	31744
layer_9	Dense	ReLU	31744	64
output	Dense	Sigmoid	64	1

Hyperparameter selection is a crucial phase in improving the proposed CNN model to ensure robust generalization when the proposed CNN model predicts depressed Arabic tweet. The specific hyperparameter configuration detailed in Table 3 is determined empirically through a systematic grid search validated on the validation set. Specifically, the adaptive moment estimation (ADAM) optimizer is selected for its adaptive learning rate properties, which accelerate convergence and handle sparse gradients efficiently in image-based textual representations. Simultaneously, binary cross-entropy is used as the standard loss function for our binary classification goal. Furthermore, the number of epochs and batch size are set to 50 and 25, respectively, to balance computational efficiency and reduce the risk of underfitting. A dropout rate of 0.5 is explicitly implemented as a regularization technique to mitigate overfitting by forcing the network to learn redundant representations. This meticulously selected suite of parameters yields the highest performance accuracy while supporting model stability. The loss function controls the loss of the proposed CNN model during training. The accuracy measure is used as a metric function. Epochs and verbose represent iterations and the method that displays details during the learning process. The rate is the fraction of the input units to drop in the Dropout layer, which takes a float value between 0 and 1.

Table 3 The hyperparameters in the proposed model

Parameter	Value
Optimizer	ADAM
Metrics function	Accuracy
Loss function	Binary Cross-Entropy
Epochs	50
Batch size	25
Random seed	32
Verbose	2
Rate	0.5

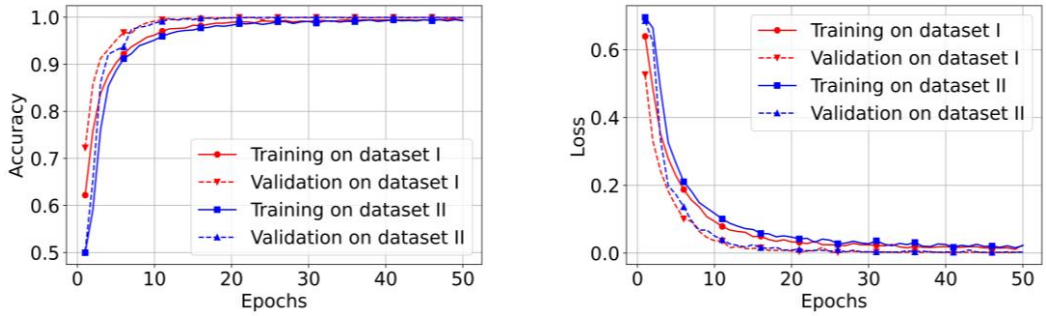
4. Results and Discussion

The proposed CNN model is built and implemented on a personal computer (PC) equipped with an AMD Ryzen 7 processor, 8 gigabytes of random access memory (RAM), and 8 gigabytes of video memory. The PC manages its resources using a 64-bit Windows 11 Home. The framework has been programmed in Python using Jupyter Notebook, which is a web-based interactive computing environment that supports multiple programming languages.

Dataset splitting is an essential methodological step in this framework that divides the generated dataset I and dataset II into discrete, non-overlapping sections to guarantee thorough model evaluation and reduce the danger of overfitting. The training set, which is used to optimize the learnable parameters of the proposed CNN model and identify the underlying patterns in the data, is allocated 60% of the total size of the dataset. The validation set is utilized to tune the hyperparameter and conduct preliminary model selection during the training process, which is 20% of the total size of the dataset. The testing set is utilized to obtain an unbiased assessment of the final model's generalization ability on unseen Arabic tweets, which is 20% of the total size of the dataset. A stratified sampling strategy is rigorously applied using the stratify parameter during the stage of the dataset split into training set, validation set, and testing set to achieve a balance in the proportion of class relative to the original dataset. These proportions are precisely maintained across all splits, ensuring a valid and unbiased evaluation framework that reinforces the mathematical rigor and reproducibility of our experimental results. This 60%, 20%, and 20% split is a standard practice that provides sufficient data for model learning, while reserving adequate portions to confirm performance and ultimately gauge the model's expected effectiveness on future real-world observations.

Fig. 8(a) illustrates the training and validation accuracy for the proposed CNN model over 50 epochs. The black line represents the training accuracy on dataset I, which shows the speed to reach one in the initial epochs, indicating that the model is effectively learning from the training data. The black dashed line corresponds to the validation accuracy on dataset I, which is higher than the training accuracy and closer to unity. Alternatively, the blue dashed line denotes the validation accuracy on dataset II, matches approximately the black dashed line after 20 epochs. While the blue line displays the training accuracy on dataset II, it increases in parallel with the black line and remains lower than the black line. The training accuracy, reaching one on the training data, indicates the model learns from the training data, decreasing the generalization.

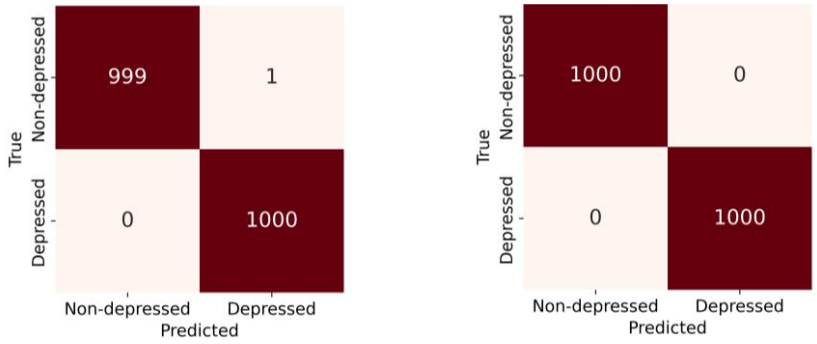
Fig. 8(b) illustrates the performance for the proposed CNN model over 50 epochs during the training process. The black line, representing the training loss on dataset I, demonstrates a rapid and monotonic decrease, stabilizing near zero after approximately 20 epochs. This shows that the proposed CNN model is effectively reducing the error on the training dataset. Conversely, the black dashed line corresponds to the validation loss on dataset I, which initially decreases in parallel with the training loss but begins to descend steadily and reaches zero after 18 epochs. On the other hand, the blue dashed line denotes the validation loss on dataset II matches approximately the black dashed line after 20 epochs. While the blue line reflects the training loss on dataset II, it decreases in parallel with the black line while staying higher than the black line. The validation loss reaching zero and the training loss not reaching zero indicate that the model has achieved a good degree of generalization to the unseen data.



(a) Training and validation accuracy (b) Training and validation loss

Fig. 8 The training results of the proposed model

The performance of the proposed CNN model is assessed using the confusion matrix, which is a square table used to evaluate the performance of the model in classification problems. The testing dataset I and dataset II consist of 2000 samples each, of which 1000 belong to the depression class and 1000 belong to the non-depression class. The proposed CNN model makes an incorrect forecast on one of the tweets, predicting it as a depressed tweet while it is a non-depressed tweet, as shown in Fig. 9(a). In contrast, the predictions of the proposed CNN model succeed entirely for dataset II, as shown in Fig. 9(b).



(a) Testing on dataset I (b) Testing on dataset II

Fig. 9 The confusion matrix of testing the proposed model

A receiver operating characteristic (ROC) curve is displayed in Fig. 10 as an additional assessment metric to examine the performance of the proposed CNN model more closely. The ability of the proposed CNN model to distinguish between classes is measured by the area under the receiver operating characteristic curve (AUROC), which is used as a summary of the ROC curve. The proposed CNN model performs better at distinguishing between the depression and non-depression classes, with a higher AUROC in both datasets I and II. Since the AUCs for the proposed CNN model in both datasets I and II are close to or equal to unity, the diagnostic test is perfect for distinguishing between depression and non-depression in Arabic tweets. In the experiments on dataset II, the proposed CNN model achieves the highest AUROC of 1, while the proposed CNN model in dataset I achieves a slightly lower AUROC of 0.999. Based on the AUROC values, the proposed CNN model in dataset II, represented by the green dashed line, is the best. A value of 1.00 indicates that the model is perfect at distinguishing between the depression and non-depression classes. Therefore, the model tested on dataset II proves to be the superior choice.

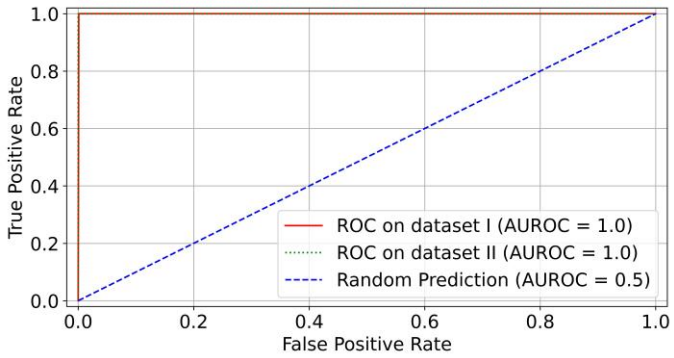


Fig. 10 The ROC curve of testing the proposed model

The proposed CNN model achieves 99.95% accuracy on the testing dataset I to detect depression in Arabic tweets. Accuracy is defined as the ratio of the correctly predicted samples to the total number of samples in the testing dataset. The accuracy of the proposed CNN model on the testing dataset II reaches 100%. The perfect metrics obtained on dataset II are not a byproduct of overfitting or artifacts, but rather reflect the specific nature of this dataset, where the informal social media texts hold highly explicit, distinct linguistic markers and explicit target keywords that allow the classifier to achieve deterministic separation. The values of F1-score, recall, precision, loss, and coefficient of determination (R2) on testing datasets I and II are in Table 4.

Table 4 The results of the proposed model

Parameter	Dataset I	Dataset II
	Value	Value
R^2	0.9980	1.0000
Recall	1.0000	1.0000
Precision	0.9990	1.0000
F1-score	0.9995	1.0000
Accuracy	0.9995	1.0000
Loss	0.0021	0.0022

The proposal, which includes generating two new datasets, uses Arabic tweets, handles punctuation, and achieves an accuracy of 100%. It showed its contribution to detecting depression in Arabic tweets. Table 5 compares the results of novel works in [28-30] to detect depression in Arabic datasets from Twitter with this framework. The metrics reported in Table 5 are cited from the original papers to provide broad context, as a direct benchmark on exactly the same dataset was not performed.

Table 5 The comparison

Ref.	F1-score	Accuracy
M. M. Alnaddaf and M. S. Basarslan [28]	0.957	0.958
H. Alkasem et al. [29]	0.940	0.949
A. Helmy et al. [30]	0.966	-
Proposal	1.000	1.000

Despite the high performance of the proposed CNN model in predicting depression in Arabic tweets, this framework has limitations. The process of converting text of Arabic tweet into images increases the computational overhead and storage requirements compared to the traditional embedding method. In contrast, while the proposed CNN model excels at capturing structural and visual patterns of text in an image, it does not explicitly account for the semantic context that advanced language models such as BERT provide. Additionally, the perfect performance of the CNN model may not fully translate to real-world, messy, or nuanced Arabic social media streams where explicit target markers are absent.

5. Conclusion

This study proposed a novel CNN model approach for detecting depression in Arabic tweets on social media by representing textual content as visual signals rather than linguistic tokens to address the challenges in the Arabic NLP field. The proposed CNN model consists of eleven layers that stack sequentially and take visual signals as input. The model was evaluated on the ArabicDeptweets_10000 dataset and achieved an accuracy rate of 100%, outperforming baseline models. These results showed the potential of visual representations to detect depression in Arabic tweets, particularly the challenges of feature extraction from text. Therefore, the proposed approach provided a potential framework for automated mental health monitoring based on Arabic social media data.

Future work will focus on improving the image-rendering process to reduce computational costs and integrating hybrid models that combine visual features with semantic embeddings for even deeper psychological profiling, using another architecture of CNN, such as VGG-16 and ResNet50, and a transformer architecture such as AraBERT.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] T. Anbesaw, M. A. Kassa, W. Yimam, A. B. Kassaw, M. Belete, A. Abera, et al., “Factors Associated with Depression Among War-Affected Population in Northeast, Ethiopia,” BMC Psychiatry, vol. 24, no. 1, article no. 376, 2024.

[2] Z. Ning, H. Hu, L. Yi, Z. Qie, A. Tolba, and X. Wang, “A Depression Detection Auxiliary Decision System Based on Multi-Modal Feature-Level Fusion of EEG and Speech,” IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 3392-3402, 2024.

[3] V. Vandana, N. Marriwala, and D. Chaudhary, “A Hybrid Model for Depression Detection Using Deep Learning,” Measurement: Sensors, vol. 25, article no. 100587, 2023.

[4] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, “Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender,” Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1248-1252, 2014.

[5] V. Tejaswini, K. S. Babu, and B. Sahoo, "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, vol. 23, no. 1, article no. 4, 2024.

[6] F. Liu, Q. Ju, Q. Zheng, and Y. Peng, "Artificial Intelligence in Mental Health: Innovations Brought by Artificial Intelligence Techniques in Stress Detection and Interventions of Building Resilience," Current Opinion in Behavioral Sciences, vol. 60, article no. 101452, 2024.

[7] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep Learning Applications and Challenges in Big Data Analytics," Journal of Big Data, vol. 2, article no. 1, 2015.

[8] S. K. Salim, M. M. Msallam, and H. Ismail, "Design Novel CNN Architecture to Protect Personal Devices from Unauthorized Access Using Face Recognition," Nanotechnology Perceptions, vol. 20, no. 3, pp. 82-93, 2024.

[9] B. Verma, S. Gupta, and L. Goel, "A Neural Network Based Hybrid Model for Depression Detection in Twitter," Communications in Computer and Information Science, vol. 1244, pp. 164-175, 2020.

[10] A. M. Helmy, R. Nassar, and N. Ramdan, "Depression Detection for Twitter Users Using Sentiment Analysis in English and Arabic Tweets," Artificial Intelligence in Medicine, vol. 147, article no. 102716, 2024.

[11] A. Leis, F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz, "Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis," Journal of Medical Internet Research, vol. 21, no. 6, article no. e14199, 2019.

[12] J. Younes, E. Souissi, H. Achour, and A. Ferchichi, "Language resources for Maghrebi Arabic dialects' NLP: A Survey," Language Resources and Evaluation, vol. 54, no. 4, pp. 1079-1142, 2020.

[13] I. Boulesnam and R. Boucetti, "Arabic Language Characteristics that Make its Automatic Processing Challenging," The International Arab Journal of Information Technology, vol. 22, no. 4, pp. 814-831, 2025.

[14] S. Lyu, X. Ren, Y. Du, and N. Zhao, "Detecting Depression of Chinese Microblog Users Via Text Analysis: Combining Linguistic Inquiry Word Count (LIWC) With Culture and Suicide Related Lexicons," Frontiers in Psychiatry, vol. 14, article no. 1121583, 2023.

[15] M. E. Villa-Pérez, L. A. Trejo, M. B. Moin, and E. Stroulia, "Extracting Mental Health Indicators from English and Spanish Social Media: A Machine Learning Approach," IEEE Access, vol. 11, pp. 128135-128152, 2023.

[16] J. Cha, S. Kim, and E. Park, "A Lexicon-Based Approach to Examine Depression Detection in Social Media: The Case of Twitter and University Community," Humanities and Social Sciences Communications, vol. 9, article no. 325, 2022.

[17] K. M. Broczek, M. C. Gely-Nargeot, and P. Gareri, "Editorial: Depression Across Cultures and Linguistic Identities," Frontiers in Psychology, vol. 15, article no. 1538489, 2024.

[18] S. Khan and S. Alqahtani, "Hybrid Machine Learning Models to Detect Signs of Depression," Multimedia Tools and Applications, vol. 83, no. 13, pp. 38819-38837, 2023.

[19] S. Zhou and M. Mohd, "Mental Health Safety and Depression Detection in Social Media Text Data: A Classification Approach Based on a Deep Learning Model," IEEE Access, vol. 13, pp. 63284-63297, 2025.

[20] A. Abdulsalam, A. Alhothali, and S. Al-Ghamdi, "Detecting Suicidality in Arabic Tweets Using Machine Learning and Deep Learning Techniques," Arabian Journal for Science and Engineering, vol. 49, no. 9, pp. 12729-12742, 2024.

[21] E. M. Rabie, A. F. Hashem, and F. K. Alsheref, "Recognition Model for Major Depressive Disorder in Arabic User-Generated Content," Beni-Suef University Journal of Basic and Applied Sciences, vol. 14, article no. 7, 2025.

[22] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Federated Learning for Privacy-Preserving Depression Detection with Multilingual Language Models in Social Media Posts," Patterns, vol. 5, no. 7, article no. 100990, 2024.

[23] M. Jogin, Mohana, M. S. Madhulika, G. D. Divya, R. K. Meghana, and S. Apoorva, "Feature Extraction Using Convolution Neural Networks (CNN) And Deep Learning," Proceedings of the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, Bangalore, India, 2018.

[24] X. Yang, J. Yan, W. Wang, S. Li, B. Hu, and J. Lin, "Brain-Inspired Models for Visual Object Recognition: An Overview," Artificial Intelligence Review, vol. 55, no. 7, pp. 5263-5311, 2022.

[25] L. Zhao and Z. Zhang, "An Improved Pooling Method for Convolutional Neural Networks," Scientific Reports, vol. 14, article no. 1589, 2024.

[26] X. Zhang, Y. Zhang, and F. Yu, "HiT-SR: Hierarchical Transformer for Efficient Image Super-Resolution," Lecture Notes in Computer Science, vol. 15098, pp. 483-500, 2025.

[27] A. Rajanand and P. Singh, "ErfReLU: Adaptive Activation Function for Deep Neural Network," Pattern Analysis and Applications, vol. 27, no. 2, article no. 68, 2024.

[28] M. M. Alnaddaf and M. S. Basarslan, "Sentiment Analysis Using Various Machine Learning Techniques on Depression Review Data," Proceedings of 8th International Artificial Intelligence and Data Processing Symposium, Malatya, Turkey, 2024.

[29] H. Alkasem, A. Alsalamah, L. Alhussan, N. Alzahrani, and S. Aba-alkhail. "Machine Learning Methods for Detecting Depression in Arabic Tweets: A Comprehensive Performance Analysis with Enhanced Evaluation Metrics." Discover Artificial Intelligence, vol. 6, article no. 120, 2026.

[30] A. Helmy, R. Nassar, and N. Ramdan. "Cross-Lingual Depression Detection for Twitter Users: A Comparative Sentiment Analysis of English and Arabic Tweets.", Preprint, 2023, doi: 10.21203/rs.3.rs-3197428/v1.