

An Enhanced Deep Learning Framework Integrating Adaptive Whale Optimization and Natural Language Processing for Mental Health Disorder Diagnosis

Nameer N. El.Emam¹, Issa Qabajeh^{2*}, Mhammed Almakadmeh³

Computer Science Department, College of Information Technology, Amman Arab University, Jordan¹

Software Engineering Department, College of Information Technology, Amman Arab University, Jordan²

École de Technologie Supérieure, Université du Québec, Montréal, Canada³

Abstract—Mental health disorders are among the biggest global public health challenges, affecting over 1 billion people worldwide and placing a burden on healthcare systems, economies, and societies. Despite progress in clinical psychiatry, mental health professionals still rely heavily on subjective judgments to make diagnoses, leading to delays and inconsistency in diagnosis-making and making early intervention difficult. The evolution of technology-enabled communication channels, i.e., social media, has enabled researchers to employ computational techniques to study digital signatures and patterns and identify psychological distress or behavioral changes in people via their electronic interactions. This study proposes a psychiatric disorder classification system based on Natural Language Processing (NLP) and an Enhanced Deep Learning model powered by the Adaptive Whale Optimization Algorithm (EDL-AWOA). It examines how users in Reddit mental health communities communicate textual data to assist a machine learning algorithm in diagnosing depression and anxiety. It also aims to build a diagnostic support tool using these findings to help identify early signs of mental illness. The results integrate existing measures into a system designed to complement clinical judgment, with the hope that clinicians will adopt these additional steps more quickly. When evaluated with standard metrics of classification accuracy, precision, recall, F1 score, and the confusion matrix, the strong results in the training phase are achieved, including 93.99% accuracy | 94.01% precision | 93.99% recall | 93.99% F1 Score.

Keywords—Psychiatric disorder classification; Natural Language Processing (NLP); Enhanced Deep Learning (E-DL); Adaptive Whale Optimization Algorithm (AWOA); reddit mental health text; depression; anxiety

I. INTRODUCTION

Psychiatric diseases, or mental disorders, are conditions that negatively affect the way you think, feel, and act; they can also hinder a person's daily functioning. They are not just "feelings"; they also shape how a person navigates social connections, whether they succeed or flounder academically or professionally, and their sense of competence in handling day-to-day challenges. In recent years, artificial Intelligence (AI), machine learning (ML), and natural language processing (NLP) have become major resources for healthcare research. Their ability to process massive amounts of unstructured text clinical notes, patient stories, and even social media posts in ways that would be practically impossible by hand is often

cited as the reason for this. Over a billion people around the world suffer from mental health disorders, such as anxiety disorders and major depressive disorder (MDD) [1]. Clinical interviews and self-report questionnaires are widely used but limited by time constraints, subjective judgment, cultural issues, and stigma. In recent studies, NLP has been used to identify psychological distress through emotional tone, sentence flow, and word choice in text [2]. Nonetheless, mental health classification is a challenging task as models are easily misled by spoken language variation, do not generalize well, and, of course, suffer from the typical training data bias [3].

To address these drawbacks, researchers have used nature-inspired optimization techniques to better guide machine learning and deep learning models. For parameter tuning, reducing overfitting, and improving classification results, researchers have applied the Whale Optimization Algorithm (WOA) in different variants [4].

Practically, the research focuses on developing an automated system that can assist in early-stage identification of mental health issues. It uses an improved deep learning approach with the Adaptive Whale Optimization Algorithm to create a robust, data-driven system that can help clinicians, psychologists, and mental health organizations detect psychiatric risk by analyzing language use. Tayebi Arasteh et al. [5] conducted research. Wongkoblap et al. [6] proposed that large language models can also facilitate more efficient, automated machine-learning workflows in clinical research. Researchers can leverage the massive amounts of data available on platforms like Reddit to detect real-time, population-level trends in mental illness, helping enable prevention efforts (2022). Researchers have also found that machine-learning models can detect and assess depression and other mental health disorders in people based on their online speech [7].

In theory, the research shows how NLP systems can leverage natural optimization to create better models of human psychological expression through language. It incorporates deep learning and the Adaptive Whale Optimization Algorithm (AWOA), which can improve a model's performance on novel data and higher-order language-independent tasks. The work connects theoretical explanations for how a representation of digital language maps across structures of human psychology to support the utility of NLP in mental health and opportunities

*Corresponding author.

for collaborative work between AI (e.g., [6]; [7]), psychology, and computational behavioral sciences.

Natural Language Processing: Natural Language Processing has come a long way, but diagnosing mental health disorders with clinical accuracy is still an open problem. Current diagnostic methods still rely heavily on clinician interpretation of the patient and their presentation, which may lead to misdiagnosis or delayed treatment. As [8] points out in their 2024 article, text-based NLP models have tried to address this by recognizing patterns in what people write. Nevertheless, they face the same limitations: model interpretability is dicey, datasets are biased, privacy and fairness are still issues, and performance drops off a cliff outside English. Most annotated datasets are aggregated in English, leaving a gap in generalization to other languages, and most models are trained for a single task but cannot perform satisfactorily enough to recognize emotional and linguistic characteristics in posts from people with mental illness. This necessitates systems that classify better, use more attention-based and contextual methods, and generalize to multilingual data.

- How effectively can an enhanced deep learning-based two-hidden-layer neural network model using the adaptive whale optimization algorithm (AWOA) classify psychiatric disorders, such as depression and anxiety, from social media posts?
- Which discourse and lexico-semantic elements in user language are the most crucial indicators for classifying particular mental health disorders?
- How does the proposed EDL-AWOA classification model compare with standard Deep Learning and WOA-based models in terms of classification performance, interpretability, and generalization?

Correspondingly, the objectives are:

- To design and implement a specialized deep learning architecture derived from the Adaptive Whale Optimization Algorithm (AWOA) for accurate classification of depression and anxiety using text data from Reddit.
- To identify and analyze the salient lexico-semantic features and structural elements of discourse in user-generated text that serve as primary indicators for distinguishing among various psychiatric conditions.
- To evaluate the optimized deep learning model against baseline deep learning and WOA-based classification models across accuracy, precision, recall, F1-score, AUC-ROC, interpretability, and generalization ability.

Finally, although the proposed EDL-AWOA framework showed promising performance in identifying mental health conditions from Reddit posts, the study has several limitations. First, the model was developed and evaluated using self-reported social media data rather than clinically verified patient records. Consequently, the system should be considered a computer-aided diagnostic support tool intended to assist with early detection and risk assessment, not to replace professional clinical evaluation. Second, the dataset was limited to English-

language Reddit content, which may limit the generalizability of the findings to other languages, cultures, and online platforms. Therefore, future research should validate the proposed framework using clinically annotated datasets and real-world healthcare data in collaboration with mental health professionals to further assess its clinical applicability, robustness, and reliability.

II. RELATED WORK

Several studies have examined machine learning, deep learning, transformer-based models, and various optimization methods for mental health classification using social media text, speech signals, and clinical datasets. In general, these works suggest that computational models can detect linguistic and behavioral cues associated with psychological distress.

Hameed et al. [9] explore the early detection of mental disorders through computational approaches. They used black-box machine learning (ML) models, including Support Vector Machines (SVM), Random Forests (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN). We employed advanced Natural Language Processing (NLP) techniques, including TF-IDF, Latent Dirichlet Allocation (LDA), N-grams, Bag of Words (BoW), and GloVe embeddings, to extract linguistic and semantic features. To address the interpretability limitations of black-box models, Explainable AI (XAI) methods were integrated, specifically Local Interpretable Model-Agnostic Explanations (LIME). However, it was based on English-language datasets and did not use advanced optimization techniques to improve model convergence and generalization.

Senanayake & Liang [10]. Transformers ensemble of DistilBERT, RoBERTa, and XLNet for depression detection in Reddit posts. This late-fusion strategy improved Macro-F1 by 6–8% over individual transformer models, suggesting that combining multiple language representations can be helpful. However, the model was resource-intensive and showed only average performance outside Reddit.

Baydili et al. In [11], a framework that uses the CW-INCA feature-selection algorithm combined with two potent deep learning models, BERT, to identify depressive and suicidal tendencies was proposed. Evaluations on six mental health datasets achieved 99.9% classification accuracy, including the Kaggle Depression Reddit dataset. Although this framework was encouraging, the model required substantial computational resources; it used only English datasets, and it did not use adaptive optimization strategies like AWOA.

Farruque et al. [12] proposed utilizing Mental-BERT along with semi-supervised learning and zero-shot learning methods to profile depression symptoms from social media text. The iterative self-training framework they proposed extends the dataset from 377 to 4,567 samples and improves the Macro-F1 score by 14%. However, several symptom classes remained underrepresented, and the dataset could not be made publicly available due to privacy limitations.

Arif et al. [13] evaluated traditional machine-learning, deep-learning, and transformer-based architectures to classify Reddit posts into six mental-illness categories. Their transformer encoder with late fusion of post titles and content

achieved the best performance, with an F1-score of 89.65%. Nonetheless, it had difficulty differentiating between disorders with the same linguistic manifestations, such as anxiety disorder and bipolar disorder, and its performance was evaluated only using Reddit data.

Nayak et al. [14] Depression Detection via TF-IDF, n-gram Features, SVM, Random Forest and LSTM Models. Their results showed that the LSTM model performed much better than traditional machine-learning methods, reaching around 85% accuracy. However, class imbalance, small dataset size, and a lack of optimization/hybrid learning mechanisms affected the study.

The work, which focused on speech-based depression diagnosis and was presented in 2024, proposed integrating Long Short-Term Memory (LSTM) networks with the Whale Optimization Algorithm (WOA). The model classified the DAIC-WOZ dataset using MFCC features, achieving about 84%. Despite the performance gain from optimization, the framework was evaluated on a single speech dataset, which is computationally expensive.

Hosseinali, F. [15] proposed an improved Selective Backpropagation algorithm (SBP+) to speed up neural network training while improving generalization. The main idea is that during training, the algorithm excludes samples that are already classified correctly with sufficient confidence, rather than applying backpropagation to all training instances at every iteration. This reduces unnecessary weight updates and focuses learning on more informative, difficult samples. However, this study introduces additional hyperparameter-tuning requirements and does not extensively investigate scalability and effectiveness across diverse large-scale application domains. Moreover, the available description does not indicate extensive validation on large-scale NLP, computer vision, or multimodal datasets.

Moosavi et al. [16] suggested using a metaheuristic algorithm, namely the Whale Optimization Algorithm (WOA), to solve constraint optimization problems. The nature-inspired algorithm follows a spiral bubble-net hunting strategy; it does not get stuck in a local minima solution even if the search space is discontinuous.

Scherbakov et al. [17]. An AI-assisted scoping review of 1,768 studies to explore the application of Natural Language Processing (NLP) techniques in mental health research and analyze how different Social Determinants of Health (SDOH) are captured within mental health analytics was conducted. The review found that most studies relied on clinical notes and social media data; few incorporated SDOH factors, confirming a significant gap in the literature. The main limitation is that this investigation is a review, not a new predictive model; other limitations include the limited use of SDOH variables in general practice, the unavailability of datasets for multiple studies described, and a lack of standardized reporting practices across the literature.

In general, past research shows that social-media text contains useful linguistic signals for detecting mental health disorders. While classification performance has been improved by deep learning, various transformer models, and better

optimization algorithms in recent years, many issues persist, such as high computational complexity, heavy dependence on the English language, dataset imbalance, limited generalization to other languages or more social science datasets like microblog texts rather than pure text forms like news or articles, etc., as discussed below. These limitations inspire the development of the EDL-AWOA model, where an advanced NLP feature extraction approach is integrated with an adaptive whale optimization technique to improve classification performance, robustness, and generalization in psychiatric disorder detection.

III. FORMAL DEFINITIONS

This section presents the formal mathematical foundations used by the proposed model, covering text feature representation, the deep learning formulation, and the whale optimization mechanism.

Definition 1: Let ID_w be the identity word level that distinguishes between various pieces of writing. The term weighting $TF_IDF_{t,d}$ of a term t in a document d is defined by the following Eq. (1):

$$ID_w : TF_{t,d} \times IDF_t \rightarrow TF_IDF_{t,d} \quad (1)$$

Where the map gives the term-frequency and inverse-document-frequency components in Eq. (2) and Eq. (3):

$$TF_{t,d} = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \quad (2)$$

$$IDF_t = \log\left(\frac{N}{df_t}\right) \quad (3)$$

where, T is the set of all terms in the corpus; D is the set of all documents; $f(t, d)$ is the number of times term t appears in document d ; N is the total number of documents; and $df(t)$ is the number of documents that contain t . $TF_IDF_{t,d}$ acts as a statistical weighting system that assigns a high value to a term that appears frequently in one document but rarely across the rest of the corpus.

Definition 2: Let the character-level (n_gram) extraction function C of a text string S be defined by the following Eq. (4):

$$n_gram: s_i \times s_{i+1} \times \dots \times s_{i+n-1} \rightarrow C_{n,S} \forall i \in [1, m - n + 1] \quad (4)$$

where, $\forall i \in [1, m - n + 1]$ represents all text strings over the alphabet Σ ; m is the string length of S ; and $P(\Sigma^n)$ is the set of character sequences of length n . n_gram is created by applying Eq. (1) to the elements of $C(n, s)$ instead of full words, allowing the model to recognize spelling variations, slang, repeated-letter patterns, and partial words.

Definition 3: Let the Latent Semantic Analysis LSA_k transformation be obtained from the truncated Singular Value Decomposition (SVD) of a term-document matrix X . A complete formal definition is typically written as the Eq. (5):

$$LSA_k: \mathbb{R}^{T \times |D|} \rightarrow \mathbb{R}^{k \times |D|} \text{ with } X \approx U_k \Sigma_k V_k^T \quad k \in |T| \quad (5)$$

where, T denotes the set of terms (words) in the vocabulary, D denotes the set of documents, X is the original term-document matrix; U_k and V_k^T contain the reduced left and right singular vectors obtained via Singular Value Decomposition; Σ_k contains the top k singular values; and k is the selected number of latent dimensions. LSA is a dimensionality-reduction technique often applied after $TF-IDF$ to reduce noise and computational cost while preserving important relationships between documents and terms [18].

Definition 4: Let a layer ℓ of a feed-forward deep learning model be defined by the following Eq. (6) and (7):

$$l: \mathbb{R}^n \rightarrow \mathbb{R}^m \text{ and } z_l = Wx + b \quad \forall x \in \mathbb{R}^n \quad (6)$$

$$a = f(z) \quad (7)$$

where, $x \in \mathbb{R}^n$ is the input vector; $W \in \mathbb{R}^{m \times n}$ is the weight matrix; $b \in \mathbb{R}^m$ is the bias vector; z is the weighted input; $f(\cdot)$ is the activation function; and a is the layer's output activation. A feed-forward deep learning model is a composition of successive layer maps that transmits information from the input layer through multiple hidden layers to the output layer [19].

Definition 5: Let the softmax output function (σ) be defined by the following Eq. (8):

$$\sigma: \mathbb{R}^K \rightarrow (0,1)^K \text{ and } \sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \forall i=1, \dots, K \quad (8)$$

where, $\sigma(z_i)$ is the predicted probability of class i ; z_i is the score for class i ; K is the total number of classes; and the denominator is the sum of exponential scores over all classes. The softmax function serves as the final layer for classification tasks, transforming output scores into probability values, and the class with the maximum probability is selected as the prediction [20].

Definition 6: Let D be the distance between the current whale location and the top (best) solution defined by the Eq. (9) and Eq. (10)-(12).

$$Distance: F(C, X^*, X_t) \rightarrow D \quad (9)$$

$$\text{where, } F(C, X^*, X_t) = |C \cdot X^* - X_t| \quad (10)$$

and

$$X_{t+1} = X^* - A \cdot D \quad (11)$$

with

$$A = 2ar - a, \quad C = 2r \quad (12)$$

where, X_t is the current solution (whale position) at iteration t ; X^* is the best solution found so far; A and C are coefficient vectors; r is a random value in $[0, 1]$; a is a control parameter that decreases linearly during the search; and $t_{\max}$ is the maximum number of iterations. The Whale Optimization Algorithm is a nature-inspired algorithm based on humpback whale hunting behavior. In this algorithm, each whale represents a candidate solution, and the algorithm balances exploration of new regions of the solution space with exploitation toward the best solution [4].

Definition 7: Let the merged labeled (MLable) dataset of D be defined by the union of the two collected Reddit mental health datasets; see the Eq. (13):

$$MLable: \bigcup_{i=1}^2 D_i \rightarrow D \quad (13)$$

where, D is defined in the Eq. (14)

$$D = \{(s_i, y_i) : s_i \in S_{raw} \text{ and } y_i \in Y\} \quad \forall i=1, \dots, n \quad (14)$$

where, D_1 and D_2 are the first and second Reddit mental health datasets after column-name standardization; s_i is the raw text of the i^{th} post; y_i is its normalized class label; S_{raw} is the set of raw Reddit posts; Y is the set of diagnostic categories; and n is the total number of samples. Records with missing text or labels are excluded from D .

Definition 8: Let S denote the set of normalized text strings over a given alphabet. Define the preprocessing function P in the Eq. (15):

$$P: REM \times EMJ \times LOW \rightarrow S \quad (15)$$

where, $LOW: S \rightarrow S$ is the lowercase transformation defined by mapping each character of a string to its lowercase equivalent. $EMJ: S \rightarrow S$ is the emoji normalization function that replaces each emoji symbol occurring in a string with its corresponding textual description according to a predefined emoji dictionary. $REM: S \rightarrow S$ is the removal function that deletes URLs, subreddit references. Thus, for any input string $s \in S$: $P(s) = REM(EMJ(LOW(s)))$, where P first converts all characters to lowercase, then replaces emoji symbols with their textual descriptions, and finally removes URLs, subreddit mentions, and user mentions.

Definition 9: Let the tokenization function TOK of a normalized string be defined by the following Eq. (16):

$$TOK: S \rightarrow W^*, \quad TOK(s) = (w_1, w_2, \dots, w_q) \quad \forall s \in S \quad (16)$$

where, W^* denotes the set of finite token sequences, and q is the number of tokens produced from the string s .

Definition 10: Let W be a vocabulary of tokens, let W^* denote the set of all finite token sequences over W , and let $SW \subseteq W$ be the set of English stop words. Define the token filtering function in Eq. (17):

$$FILT: W^* \rightarrow W^* \text{ And } FILT(w_1, \dots, w_q) = (w_i : w_i \notin SW \text{ and } |w_i| > 1) \quad (17)$$

where, $|w_i|$ denotes the number of characters in token w_i .

Thus, $FILT$ removes all stop words and all tokens of length one while preserving the original relative order of the remaining tokens.

Definition 11: Let the WordNet lemmatization function LEM be defined by the following Eq. (18) to replace each token with its lemma:

$$LEM: W \rightarrow \Lambda, \quad LEM(w_1, \dots, w_q) = (LEM(w_1), \dots, LEM(w_q)) \quad (18)$$

where, $\Lambda \subseteq W$ is the set of WordNet lemmas (canonical dictionary forms), so each inflected token maps to its base form.

Definition 12: Let S_{raw} denote the set of raw text strings and let S_{clean} denote the set of cleaned token sequences in Eq. (19).

$$\text{CLEAN}: S_{\text{raw}} \rightarrow S_{\text{clean}} \text{ and}$$

$$\text{CLEAN} = \text{REC} \circ \text{LEM} \circ \text{FILT} \circ \text{TOK} \circ \text{NORM} \quad (19)$$

where, $\text{REC}: W^* \rightarrow S_{\text{clean}}$ reconstructs a cleaned text sequence from the resulting lemmatized tokens.

Hence, for every raw text string $s \in S_{\text{raw}}$, and $\text{CLEAN}(s) = \text{REC}(\text{LEM}(\text{FILT}(\text{TOK}(\text{NORM}(s))))))$

Thus, CLEAN transforms a raw text string into its fully normalized, tokenized, filtered, lemmatized, and reconstructed cleaned representation.

Definition 13: Let $m \in \mathbb{N}$ be the dimension of a hidden-layer activation space, and let $p \in (0, 1)$ denote the dropout rate; the dropout regularization function DROP_p is defined in Eq. (20):

$$\text{DROP}_p: \mathbb{R}^m \rightarrow \mathbb{R}^m \text{ and } \text{DROP}_p(a) = \frac{a \odot m}{1-p} \quad (20)$$

$$\text{where, } m = (m_1, \dots, m_m)$$

$\in \{0,1\}^m$ and m_i is a sample of $\text{Bernoulli}(1 - p) \forall i = 1, \dots, m$

Here, $a = (a_1, \dots, a_m) \in \mathbb{R}^m$ is the activation vector of a hidden layer, m is a binary dropout mask, $\odot$ denotes elementwise multiplication, and p is the probability that a neuron is dropped.

The scaling factor $1/(1 - p)$ implements inverted dropout, ensuring that $E[\text{DROP}_p(a)] = a$, so that the expected activation remains unchanged during training. During training, DROP_p randomly deactivates a fraction p of the neurons to reduce overfitting and improve generalization. During inference, dropout is disabled, and the mapping reduces to the identity function $\text{DROP}_p(a) = a$. Thus, DROP_p acts as a stochastic regularization operator that randomly masks hidden-unit activations while preserving their expected magnitude.

Definition 14: Let the categorical cross-entropy loss function L and the gradient-based weight-update map be defined by the following Eq. (21) and (22):

$$L: (0,1)^K \times Y \rightarrow \mathbb{R}^+ \text{ and } L(\sigma(z), y) = - \sum_{j=1}^K 1[y=j] \log \sigma(z)_j \quad (21)$$

where, $\sigma(z)$ is the softmax output of the proposed deep learning (see Definition 5); $y \in Y$ is the true class label; $1[\cdot]$ is the indicator function.

Furthermore, the cross-entropy loss function L is applied to update the weight (w_{ij}) connecting neuron i to neuron j , using the learning rate (α) and the gradient of the loss function $\nabla L(\cdot)$ for the previous weight [21]. The training process is based on mini-batch backpropagation, Eq. (20), to calculate the new weight w_{ij}^{new} .

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \alpha \nabla L(w_{ij}^{\text{old}}) \quad \forall i, j \quad (22)$$

IV. THE PROPOSED EDL-AWOA MODEL

A. EDL-AWOA Architecture

The proposed EDL-AWOA model uses a hybrid machine learning setup that pairs deep learning with an adaptive whale optimization method. It begins with raw data collection, then moves to preprocessing for data cleaning and transformation. The processed data feeds into an enhanced deep learning design with an input layer, several hidden layers, and an output layer that produces initial predictions. At the same time, an adaptive whale optimizer improves the model by tuning parameters through adaptive search and refinement strategies inspired by whale behavior. The interaction between the deep learning model and the optimizer improves learning performance, and the final outputs are passed to the classification and evaluation stage, where the model's predictive performance is assessed. Fig. 1 shows this full workflow.

The Enhanced MLP identifies psychiatric conditions from linguistic representations of Reddit posts. The input layer processes TF-IDF, char-TF-IDF, or LSA feature vectors. The model uses two fully connected hidden layers, each with ReLU activation followed by dropout regularization. Two layers were chosen to keep the design simple and effective while still learning more complex nonlinear patterns from the extracted text features. Dropout was added to reduce overfitting, helping the system generalize better to unseen Reddit posts. The output layer applies a softmax function and produces a probability distribution over six target classes: Anxiety/Stress, Bipolar, Major Depressive, Normal, Personality Disorder, and Substance Abuse. In addition, Fig. 1 shows the enhanced deep learning architecture.

For feature extraction, the study uses four techniques: word-level TF-IDF, character-level TF-IDF, a combination of word- and character-level TF-IDF, and Latent Semantic Analysis with TF-IDF using Truncated SVD for dimensionality reduction. Dimensionality reduction is applied in specific situations to reduce operational costs and maintain proper memory performance in deep learning.

The Adaptive Whale Optimization Algorithm (AWOA) tunes the main hyperparameters of the enhanced deep learning model. In this implementation, the optimized hyperparameters are the learning rate and the dropout rate, and each whale agent represents a candidate hyperparameter configuration. The learning rate is searched in the range $[10^{-4}, 10^{-2}]$, and the dropout rate in the range $[0.10, 0.50]$. AWOA improves the standard Whale Optimization Algorithm by combining opposition-based learning, Lévy flight exploration, adaptive inertia weight, adaptive spiral movement, and an epoch-switched alpha mechanism. The epoch-switched alpha mechanism changes the alpha value within the same optimization run based on the training epoch phase: the first alpha value applies during the early training phase, before the switch epoch, while the second alpha value applies during the later training phase, after the switch epoch. For example, in the schedule $0.7 \leftrightarrow 0.3$, the optimizer uses 0.7 before the switch epoch and 0.3 afterward. The reverse schedule, $0.3 \leftrightarrow 0.7$, is evaluated separately to study how the order of alpha values

affects exploration–exploitation behavior and classification performance.

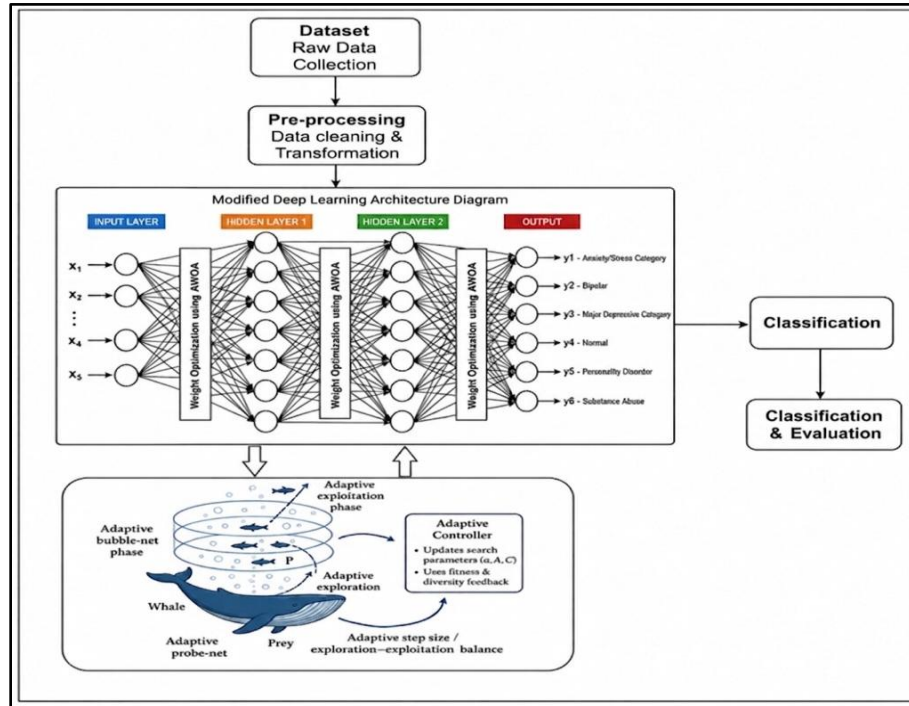


Fig. 1. The architecture of the EDL-AWOA model.

B. Algorithms of the Proposed Model

The methodology is implemented using three algorithms: a data pre-processing algorithm (Algorithm 1), the enhanced deep learning algorithm (Algorithm 2), and the adaptive whale optimization algorithm (Algorithm 3).

Algorithm 1: Data Pre-Processing Algorithm

Input: Raw Reddit text dataset

Output: Cleaned text dataset

Procedure:

- Step 1: Load and merge the collected datasets. (see Definition 7)
- Step 1.1: Load the first Reddit mental health dataset
 - Step 1.2: Load the second Reddit mental health dataset.
 - Step 1.3: Standardize the column names in both datasets, such as text and label.
 - Step 1.4: Append the rows of the second dataset to the first dataset.
 - Step 1.5: Create one unified dataset containing all Reddit text samples and their corresponding labels.
- Step 2: Normalize class labels. (see Definition 7)
- Step 3: Remove records with missing text or labels. (see Definition 7)
- Step 4: For each text sample: (see Definition 8)
- Step 4.1: Convert text to lowercase. (see Definition 8, map LOW)
 - Step 4.2: Convert emojis to textual descriptions. (see Definition 8)
 - Step 4.3: Remove URLs, Reddit mentions, and user mentions. (see Definition 8)
- Step 5: Tokenize the cleaned text. (see Definition 9)
- Step 6: Remove stop words and one-character tokens. (see Definition

10)

Step 7: Apply WordNet lemmatization. (see Definition 11)

Step 8: Reconstruct the cleaned text. (see Definition 12)

Step 9: Remove posts with fewer than three words. (see Definition 12)

Step 10: Remove duplicate posts. (see Definition 12)

Step 11: Return the cleaned dataset. (see Definition 12)

Step 12: End.

Algorithm 1 integrates two Reddit mental health datasets into a unified corpus by standardizing column names, merging records, normalizing class labels, and removing instances with missing values to ensure data consistency and quality. It then applies a sequence of text-cleaning operations, including lowercasing, emoji-to-text conversion, removal of URLs and user mentions, tokenization, stop-word filtering, and WordNet lemmatization, thereby reducing noise and standardizing linguistic variations. Finally, the cleaned tokens are reconstructed into text, while very short posts and duplicate entries are discarded, resulting in a high-quality dataset suitable for reliable machine learning and natural language processing tasks.

We first merged the Reddit datasets after standardizing class labels and column names, then removed records with missing values. Text preprocessing included lowercasing, emoji-to-text conversion, removal of URLs, subreddit references, and user mentions, followed by tokenization, stop-word removal, elimination of single-character tokens, and WordNet lemmatization. We extracted features using word-level TF-IDF (n-gram range = 1–2, maximum vocabulary = 20,000), character-level TF-IDF (n-gram range = 3–5,

maximum vocabulary = 10,000), and LSA with 300 latent dimensions via Truncated SVD. The E-DL model employed two hidden layers with 128 and 64 neurons using ReLU activation, while the dropout rate (0.10–0.50) and learning rate (10^{-4} – 10^{-2}) were optimized by AWOA; training was conducted with a batch size of 32, 32 epochs, an AWOA population size of 30 whales, and 15 optimization iterations.

Algorithm 2: Enhanced Deep Learning Model (E-DL)

Input: TF-IDF feature vectors; Class labels

Output: Predicted class; Evaluation results

Procedure:

Step 1: Load the pre-processed dataset from Algorithm 1. (see Definition 12)

Step 2: Convert the text data into TF-IDF feature vectors. (see Definition 1-3)

Step 3: Define the Enhanced DL architecture (see Definition 4)

Step 3.1: Define the input layer according to the TF-IDF feature size. (see Definition 4, input $x \in \mathbb{R}^d$)

Step 3.2: Define the first fully connected dense layer with 128 neurons and ReLU activation. (see Definition 4)

Step 3.3: Apply dropout regularization. (see Definition 13)

Step 3.4: Define the second fully connected dense layer with 64 neurons and ReLU activation. (see Definition 4)

Step 3.5: Apply dropout regularization. (see Definition 13)

Step 3.6: Define the softmax output layer for class prediction. (see Definition 5)

Step 4: Define the loss function and evaluation metric. (see Definitions 14 and 16)

Step 5: Train the E-DL using mini-batch learning and backpropagation. (see Definition 14)

Step 6: Optimize the E-DL hyperparameters using AWOA, as described in Algorithm 3, until the maximum number of epochs is reached. (see Definition 6)

Step 7: Evaluate the trained model using accuracy, precision, recall, F1-score, and AUC. (see Definition 16)

Step 8: Save the final results.

Step 9: End.

The proposed Enhanced Deep Learning (E-DL) algorithm transforms the preprocessed Reddit text into TF-IDF feature vectors and processes them through a multi-layer neural network with two fully connected, ReLU-activated dense layers (128 and 64 neurons) and dropout regularization to improve feature learning and prevent overfitting. The model is trained using mini-batch backpropagation, while the Adaptive Whale Optimization Algorithm (AWOA) optimizes hyperparameters to improve classification performance. Finally, the trained E-DL model predicts mental health categories and is evaluated using comprehensive metrics, including accuracy, precision, recall, F1-score, and AUC, to measure its effectiveness and robustness.

Algorithm 3: Adaptive Whale Optimization Algorithm (AWOA)

Input: deep learning architecture defined in Algorithm 2 and in Fig. 1.

Set the Initial value of the learning rate search space $\alpha \in [10^{-4}, 10^{-2}]$

Where 10^{-4} is the lower bound and 10^{-2} is the upper bound.

Output: Find the best α

Procedure

Step 1: Define the initial set of weight functions w_{ij} in deep learning. //see Definition 4

Step 2: Set the initial whale position (X_0) and set the learning rate (α) as in Eq. (23) //see Definition 6

$$\alpha = X_0 \quad (23)$$

Step 3: For $s = 0$ to $N - 1$.

where, $N - 1$ is the last whale position

Step 3.1: Calculate the new Whale Position X_{s+1} using Eq. (24) //see Definition 6

$$X_{s+1} = X_s - \bar{F}(X_s, X_p) \quad (24)$$

where, X_s is the s^{th} position of whales and X_p is the prey location, and $\bar{F}(\cdot)$ is the whale optimization model [22].

Step 3.2: Apply the Adaptive WOA approach based on the fitness function $Fit(\cdot)$

to calculate the best learning rate α as in Eq. (25) and Eq. (26):

$$\alpha = \beta X_{s+1} + (1 - \beta) X_s \quad (25)$$

where

$$\beta = \begin{cases} 0.9, & \text{if } \text{Fit}(X_{s+1}) > \text{Fit}(X_s) \\ 0.1, & \text{if } \text{Fit}(X_{s+1}) \leq \text{Fit}(X_s) \end{cases} \quad (26)$$

Step 3.3: Calculate the new value of weight w_{ij}^{new} using Eq. 22 //see Definition 14,

Step 3.4: Apply the training process of deep learning using an activation function between every two layers using Eq. (27): (see Definition 4, maps 6–7)

$$z = f(b + \sum_i \sum_j a_{i,j} \times w_{ij}^{\text{new}}) \quad (27)$$

where, $f(\cdot)$ is the ReLU activation function, and b is the DL bias.

Step 4: End // Find the best α of DL

The Adaptive Whale Optimization Algorithm (AWOA) is integrated with the Enhanced Deep Learning (E-DL) model to dynamically optimize the learning rate during training by simulating whale-hunting behavior. At each iteration, the WOA search mechanism updates whale positions, and a fitness function evaluates candidate learning rates to select the value that minimizes model loss and improves convergence. The selected learning rate then updates network weights via backpropagation, while ReLU activation functions enable effective feature learning across layers, yielding an optimized deep learning model with improved classification performance.

The AWOA algorithm optimizes classification accuracy on the validation dataset, defined as (Validation accuracy, Validation F1-score, Validation Macro AUC). We repeated each optimization experiment more than 20 times to minimize stochastic effects and recorded the average performance. We fixed the best hyperparameter configuration from the validation set and evaluated it on the independent test set without further optimization to prevent data leakage.

V. RESULTS AND DISCUSSION

This chapter discusses the experimental outcomes from implementing the proposed EDL-AWOA model. The main goal is to test how well the proposed model can classify psychiatric disorders using Reddit mental health text data, with

a special focus on depression and anxiety. We evaluate the results using common classification metrics, including accuracy, precision, recall, F1-score, macro AUC, and confusion matrix analysis, and compare them with baseline machine learning and deep learning models to assess the contribution of the AWOA optimization process.

A. System Setup

The research was conducted on a Colab server with 64GB of RAM, running the G4 runtime with an NVIDIA RTX PRO 6000 GPU and approximately 96GB of VRAM, with CUDA 13.0-accelerated training. Python 3.11 (managed via pyenv) was the primary programming language. PyTorch was used for model building and training; Scikit-learn for preprocessing, feature extraction, evaluation, and basic machine learning methods; NLTK for natural language preprocessing; and imbalanced-learn for handling class imbalance. NumPy and Pandas were used for data analysis, and Matplotlib for visualizing model performance. The system supports repeatable experiments by automatically logging model configurations, confusion matrix outputs, hyperparameter settings, and the final trained model weights.

The dataset is the public Kaggle Reddit Mental Health Dataset, which contains posts from mental health-related subreddits focused on depression and anxiety. We retained only English-language posts with legitimate self-reported mental health expressions. We split the dataset into 70% training, 15% validation, and 15% testing sets, and Table I shows the training and optimization configuration.

TABLE I. TRAINING AND OPTIMIZATION PARAMETERS

Parameter	Value/Condition
Batch size	32
Learning rate	0.001 default; optimized from 0.0001 to 0.01
Epochs	32
Pre-training Epochs	4
Noise dimension	100
Hidden layer 1	Dense, 128 neurons, ReLU
Hidden layer 2	Dense, 64 neurons, ReLU
Hyperparameters optimized	Learning rate, Dropout
AWOA – Population Size	30
AWOA – Max Iterations	15

The batch size was set to 32 to balance training stability and computational efficiency. The default learning rate was fixed at 0.001 for steadier convergence but was set within the AWOA search bounds of 0.0001 to 0.01. The number of training epochs was set to 32 to give the model enough opportunity to learn meaningful patterns without excessive compute time. Two dense hidden layers were used: the first with 128 neurons and ReLU activation, and the second with 64 neurons and ReLU activation. We optimized the learning rate and dropout, since both strongly affect performance. For AWOA, the population size was set to 30 whale agents and the maximum number of iterations to 15, providing a workable balance between optimization quality and computational cost.

B. Experiments on the Proposed Model to Find the Best Parameter Tuning

Let the evaluation metrics be defined by the following maps (28–32) from the confusion-matrix counts of the test set:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (28)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (29)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (30)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (31)$$

$$\text{AUC} = \int_0^1 TPR d(FPR) \quad (32)$$

where, TP, TN, FP, and FN are the true-positive, true-negative, false-positive, and false-negative counts, respectively; TPR = Recall is the true-positive rate; and FPR = FP / (FP + TN) is the false-positive rate. For the multi-class setting, the metrics are macro-averaged over the K diagnostic categories.

The proposed model was evaluated across different conditional alpha (α) schedules during training. The results show that model performance varied according to the selected α setting, indicating that the adaptive coefficient directly affects the optimization process and classification performance. Among the tested settings, Experiment 7, using the α schedule $0.9 \leftrightarrow 0.1$, achieved the best performance, with 93.99% accuracy, 94.01% precision, 93.99% recall, 93.99% F1-score, and 0.9993 macro AUC. Table II presents the full results.

TABLE II. EDL-AWOA PERFORMANCE ACROSS ALPHA SETTINGS (TRAINING)

Exp.	Adaptive α Selection	Accuracy	Precision	Recall	F1-Score	Macro AUC	Adaptive α Selection	Accuracy
1	$0.7 \leftrightarrow 0.3$	87.20%	87.22%	87.20%	87.19%	0.9968	1	$0.7 \leftrightarrow 0.3$
2	$0.3 \leftrightarrow 0.7$	85.80%	85.81%	85.80%	85.79%	0.9959	2	$0.3 \leftrightarrow 0.7$
3	$0.8 \leftrightarrow 0.2$	89.51%	89.56%	89.51%	89.51%	0.9978	3	$0.8 \leftrightarrow 0.2$
4	$0.2 \leftrightarrow 0.8$	86.58%	86.64%	86.58%	86.59%	0.9964	4	$0.2 \leftrightarrow 0.8$
5	$0.6 \leftrightarrow 0.4$	88.40%	88.43%	88.40%	88.40%	0.9973	5	$0.6 \leftrightarrow 0.4$
6	$0.4 \leftrightarrow 0.6$	87.61%	87.62%	87.61%	87.61%	0.9969	6	$0.4 \leftrightarrow 0.6$
7	$0.9 \leftrightarrow 0.1$	93.99%	94.01%	93.99%	93.99%	0.9993	7	$0.9 \leftrightarrow 0.1$
8	$0.1 \leftrightarrow 0.9$	86.21%	86.23%	86.21%	86.21%	0.9962	8	$0.1 \leftrightarrow 0.9$

The accuracy was between 85.80% and 93.99% in the eight experiments. Experiment 7 achieved the highest accuracy using the $0.9 \leftrightarrow 0.1$ schedule, indicating that the best balance between exploration and exploitation came from starting with a larger alpha value and selecting lower alpha values after the switch epoch. Similarly, the schedule from step $0.3 \leftrightarrow 0.7$ (the lowest accuracy) achieved 85.80% accuracy.

C. Confusion Matrix and AUC-ROC Curves for EDL-AWOA Model

The confusion matrix corresponding to the best EDL-AWOA configuration is shown in Fig. 2.

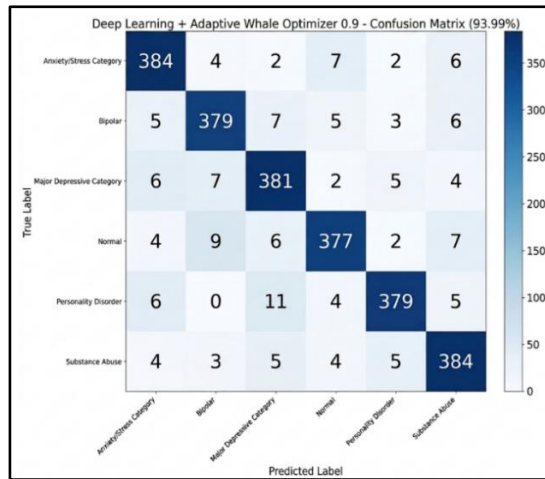


Fig. 2. Confusion matrix for the EDL-AWOA model.

The confusion matrix shows excellent classification performance with an overall accuracy of 93.99%, as indicated by the strong diagonal values (377-384 correct predictions per class) and relatively few misclassifications. Most errors occur between clinically similar classes, particularly Personality Disorder and Major Depressive Category (11 cases), and Normal and Bipolar (9 cases), suggesting some overlap in the learned features. Overall, the Deep Learning model optimized with the Adaptive Whale Optimizer (0.9) demonstrates high reliability and balanced performance across all six classes, with no class exhibiting substantially higher error rates than the others.

Fig. 3 shows the AUC-ROC curves across the six mental health categories shown in Fig. 2.

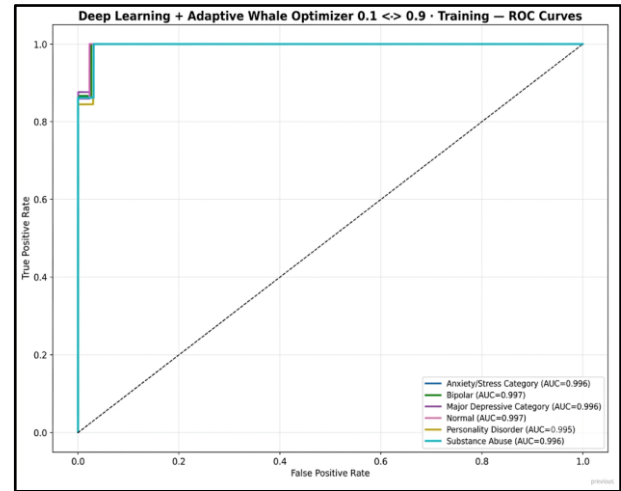


Fig. 3. AUC-ROC curves for the best EDL-AWOA model.

The AUC-ROC curves of the best EDL-AWOA model demonstrate excellent discriminative capability across all six mental health categories, with AUC values ranging from 0.995 to 0.997. The Bipolar and Normal classes achieved the highest AUC of 0.997, while Anxiety/Stress, Major Depressive Category, and Substance Abuse attained 0.996, and Personality Disorder reached 0.995. The ROC curves lie very close to the upper-left corner, indicating a high true positive rate and a very low false positive rate. These results confirm that the EDL-AWOA model provides highly accurate and robust class separation, making it highly effective for multi-class mental health disorder classification.

D. Comparison Between Various Types of Deep Learning

The proposed model (EDL-AWOA) is compared with two baselines: standard deep learning (DL) and deep learning with the standard Whale Optimization Algorithm (DL-WOA).

Fig. 4 shows the accuracy and loss curves for the training, validation, and testing phases of the Deep Learning (DL) model.

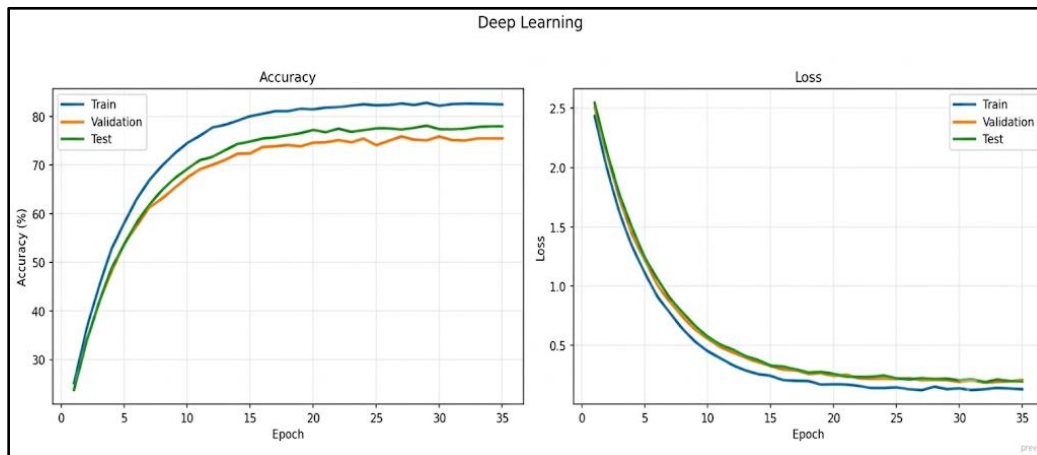


Fig. 4. Training, validation, and testing accuracy/loss curves for the Deep Learning (DL) model.

The deep learning model showed substantial progress, with training accuracy rising from about 25% to 83% (+232%), validation accuracy from 24% to 75% (+213%), and test accuracy from 24% to 78% (+225%) over 35 epochs. At the same time, the loss decreased markedly, with training loss dropping from about 2.45 to 0.12 (95.1% reduction), validation loss from 2.50 to 0.20 (92.0% reduction), and test loss from

2.55 to 0.21 (91.8% reduction). These results indicate effective model convergence and strong generalization, as the validation and test curves closely follow the training curves with only a small performance gap.

Fig. 5 shows the accuracy and loss curves for the training, validation, and testing phases of the Deep Learning (DL) with WOA (DL-WOA) model.

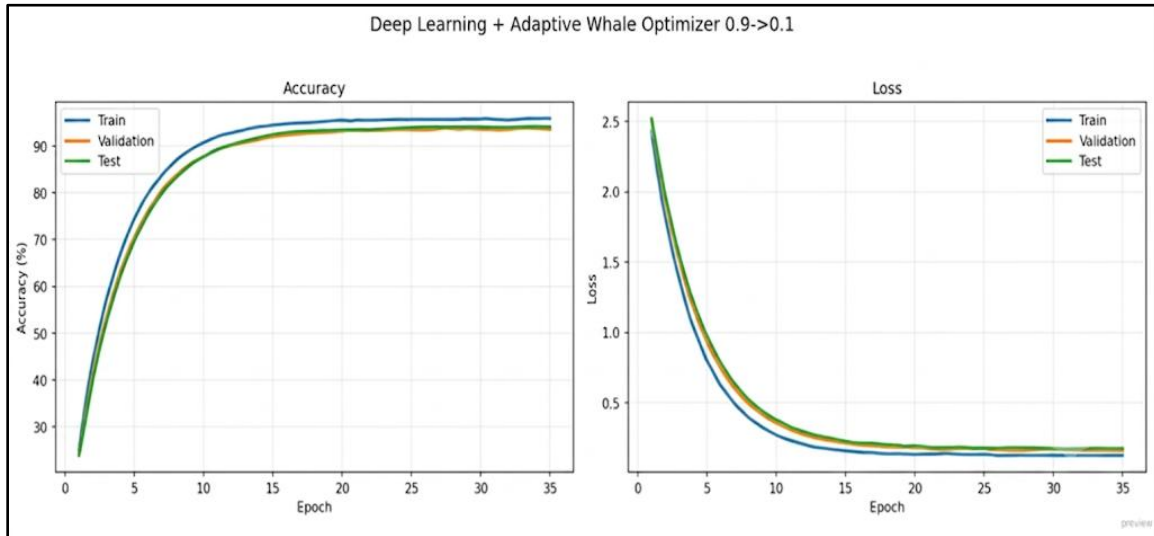


Fig. 5. Training, validation, and testing accuracy/loss curves for the Enhanced Deep Learning and adaptive WOA (EDL-WOA) model.

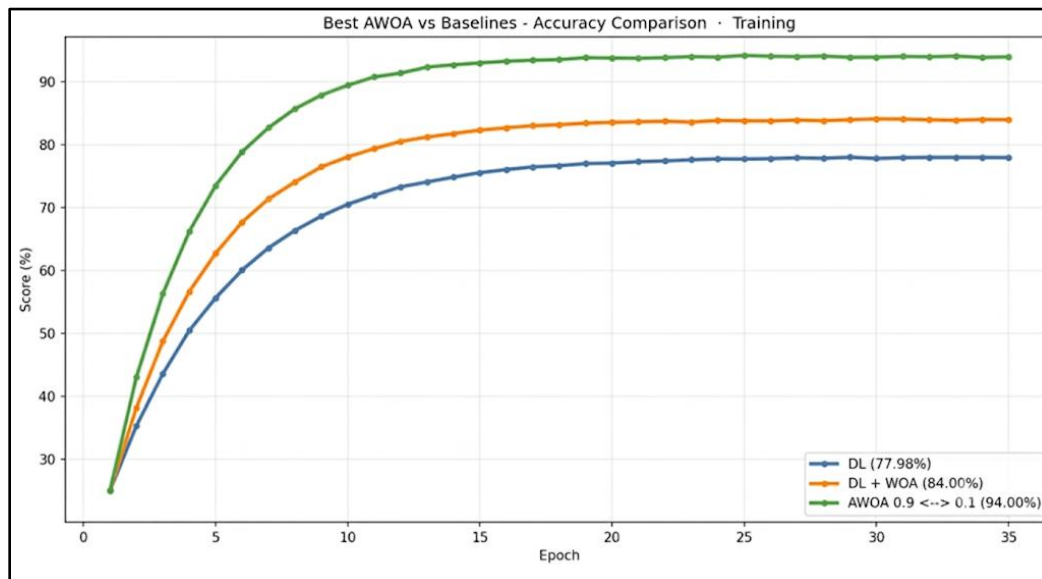


Fig. 6. Accuracy comparison between DL, DL-WOA, and EDL-AWOA (training).

The Deep Learning + Whale Optimizer model outperformed the baseline model, with training, validation, and test accuracies rising from approximately 25% to 87% (+248%), 24% to 82% (+242%), and 24% to 85% (+254%), respectively. Simultaneously, training, validation, and test losses fell from about 2.45, 2.50, and 2.55 to 0.13, 0.18, and 0.20, representing reductions of approximately 94.7%, 92.8%, and 92.2%, respectively, indicating faster convergence, improved generalization, and better predictive performance compared with the standard deep learning model.

Fig. 6 shows the training, validation, and test curves for the best EDL-AWOA configuration using the $0.9 \leftrightarrow 0.1$ alpha schedule. The accuracy curves increase rapidly during the early epochs and then stabilize after approximately epoch 15, while the loss curves decrease sharply and remain low across training, validation, and test data. The small gap between the curves indicates stable generalization and no major overfitting.

The accuracy and loss curves in Fig. 8 indicate that the proposed deep learning integrated with Adaptive Whale

Optimization (EDL-AWOA) converges quickly and achieves stable performance on the training, validation, and test datasets. Accuracy rises quickly from around 25% (epoch 1) to 95%, 94%, and 93.5% on the training, validation, and test sets, respectively. At the same time, the loss drops from 2.5 to 0.15 on training, 2.5 to 0.18 on validation, and 2.5 to 0.17 on testing, a decrease of around 92%- 94%. The closely aligned

training, validation, and test curves indicate strong generalization with little overfitting. Overall, the results show that AWOA effectively optimizes the learning process by speeding convergence, improving classification accuracy, and significantly reducing prediction error.

The accuracy comparison on the training and testing sets is shown in Fig. 7 and 8, respectively.

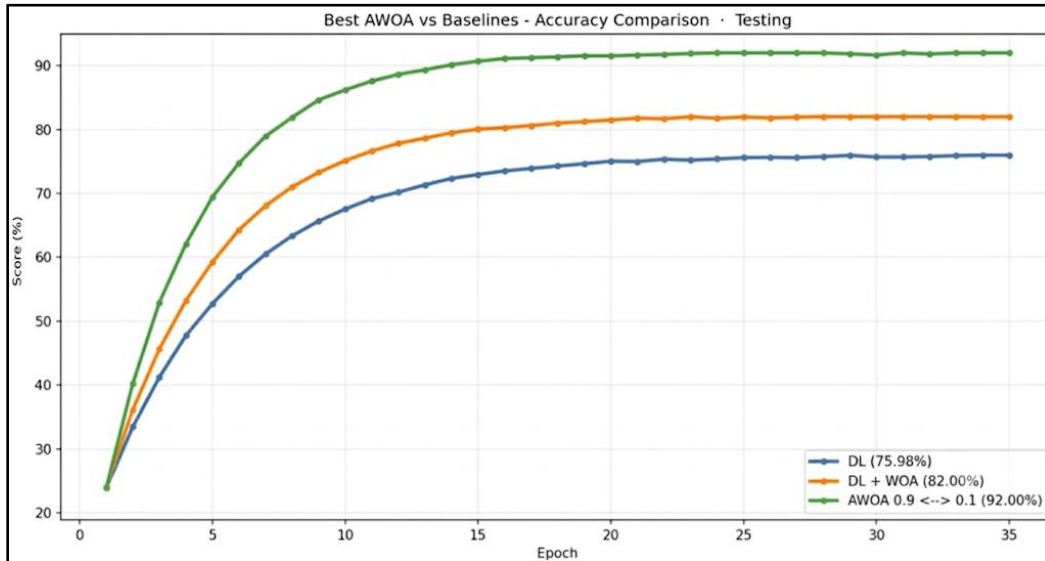


Fig. 7. Accuracy comparison between DL, DL-WOA, and EDL-AWOA (testing).

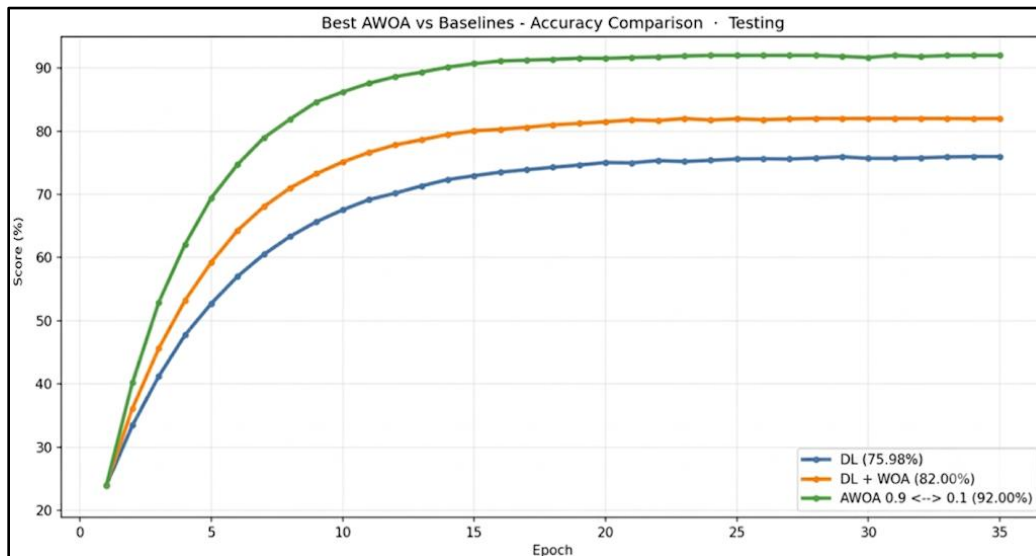


Fig. 8. Accuracy comparison between DL and DL-AWOA (testing).

The training results indicate that the proposed EDL-AWOA model outperforms both conventional DL and DL-WOA. The deep learning model (baseline DL) reached a training accuracy of 77.98%, and incorporating the Whale Optimization Algorithm (DL-WOA) yielded 84%, a 7.72% improvement over DL. The proposed EDL-AWOA model further improved training accuracy to 94.00%, an 11.90% improvement over DL-WOA and a substantial 20.54% improvement over the baseline DL model. In addition, EDL-AWOA converged faster and maintained the highest accuracy throughout training,

highlighting the effectiveness of the adaptive optimization mechanism in improving parameter tuning, feature learning, and overall classification performance.

The testing results confirm that the proposed EDL-AWOA model significantly outperforms both the baseline DL and the DL-WOA models in classification accuracy. The conventional DL model achieved a testing accuracy of 75.98%, while incorporating the Whale Optimization Algorithm increased accuracy to 82.00%, a 7.92% improvement over DL. The

proposed EDL-AWOA model further improved testing accuracy to 92.00%, an additional 12.20% improvement over DL-WOA and a remarkable 21.08% improvement over the baseline DL model. Furthermore, EDL-AWOA converged faster and maintained consistently higher accuracy across all testing epochs, showing that the adaptive optimization strategy

improves model generalization, parameter tuning, and predictive performance on unseen data.

Tables III and IV compare the proposed model (EDL-AWOA) with two baselines on the metrics (Accuracy, Precision, Recall, F1-score, and AUC) for training and testing, respectively.

TABLE III. COMPARISON OF THE PROPOSED MODEL WITH BASELINE MODELS (TRAINING)

Model	Accuracy	Precision	Recall	F1-Score	Macro AUC
DL	77.98%	77.98%	77.98%	77.98%	0.9903
DL-WOA	84%	84.05%	84%	83.97%	0.9949
EDL-AWOA (Proposed)	94%	94.01%	94%	94%	0.9993

As evaluated against the ground truth, Table III indicates that the performance of each evaluation metric remains strong on the training side, with the highest scores by EDL-AWOA. For accuracy, EDL-AWOA was 94.00%, which is better than both baseline DL (77.98%) and DL-WOA (84.00%), with improvements of 20.54% and 11.90%, respectively. The same trend is also observed with respect to precision, in which case EDL-AWOA had a performance of 94.01%, while the corresponding DL and DL-WOA were 77.98% and 84.05%. Moreover, the recall and F1-score values also improved from 77.98%(DL), approximately 84.00% (DL-WOA), to 94.00% with EDL-AWOA, respectively, which demonstrate the

superior classification consistency and balanced performance of the EDL-AWOA method. Moreover, it obtained the highest Macro AUC (0.9993) among all experimental models, including DL (0.9903) and DL-WOA (0.9949), indicating that the classification characteristics of mental disorders were clearly distinguished by the proposed model from each other. The results indicate that integrating the Adaptive Whale Optimization Algorithm could effectively improve performance in feature extraction, parameter tuning, and classification accuracy compared to its non-AWOM counterparts.

TABLE IV. COMPARISON OF THE PROPOSED MODEL WITH BASELINE MODELS (TESTING)

Model	Accuracy	Precision	Recall	F1-Score	Macro AUC
DL	75.98%	75.98%	75.98%	75.98%	0.9853
DL-WOA	82%	82.05%	82%	81.97%	0.9899
EDL-AWOA (Proposed)	92%	92.01%	92%	92%	0.9990

Table IV clearly shows the testing performance of the baseline DL and DL-WOA models against the proposed model (EDL-AWOA), with EDL-AWOA greatly outperforming them on every evaluation metric. The accuracy of the conventional DL model was 75.98%, while with the Whale Optimization Algorithm, it improved to 82.00%, i.e., a gain of 7.92% over DL. The proposed EDL-AWOA model significantly improved the accuracy to 92.00%, a relative increase of 12.20% over DL-WOA and an astounding 21.08% over the baseline DL model, respectively. As shown in Table III, a similar performance improvement is observed in precision, recall, and F1-score; EDL-AWOA obtained an accuracy of about 92.00%, while DL was about 75.98% and DL-WOA was approximately 82.00%, which indicates that EDL-AWOA has better predictive reliability than other methods with better class balance. Additionally, EDL-AWOA achieved the highest Macro AUC

of 0.9990 in this test, again higher than DL (0.9853) and DL-WOA (0.9899), demonstrating its superior ability to classify mental health categories with one-vs-all comparisons. These results indicate that the proposed adaptive whale optimization strategy improves model generalization and classification performance on unseen test data.

E. Comparison with Previous Works

Table V compares the proposed EDL-AWOA model with recent studies evaluated on the same Kaggle Reddit mental health dataset. Using a common corpus minimizes the influence of dataset-specific factors and provides a more reliable assessment of relative model performance. The results demonstrate that EDL-AWOA achieves superior classification performance compared with existing approaches evaluated under the same data conditions.

TABLE V. COMPARISON OF THE PROPOSED MODEL WITH RELATED STUDIES

Study	Method	Dataset	ACC.	Prec.	Recall	F1	AUC
Utsa et al. [23].	Interpretable linguistic features (author-disjoint)	RMHD	89%	93%	85%	89%	0.9500
Jena et al. [24].	Continual Neuro-Symbolic Graph Learning (MH-Freeze)		92%	N/R	N/R	92%	0.9420
Kim et al. [25].	Random Forest (Text + LIWC), binary depression		91%	96%	90%	93%	N/R
Proposed Model	EDL-AWOA		93.99%	94.01%	93.99%	93.99%	0.9993

The comparative study shows that the EDL-AWOA model achieved the best performance among RMHD-based studies on mental health classification. Utsa et al. (2026) on the RMHD dataset achieved 89% accuracy and an AUC of 0.9500, and Jena et al. (2026) achieved 92% accuracy and a 92% F1-score, while the proposed model performed better, with maximum accuracy (93.99%), recall (93.99%), and AUC (0.9993). Compared with Kim et al., a Random Forest classifier with text and LIWC features achieved 91% accuracy and a 93% F1-score, while the EDL-AWOA improves accuracy by c.a. 3.29% and the F1-score by 1.06%. Moreover, the exceptionally high AUC of 0.9993 indicates near-perfect class discrimination, substantially exceeding AUC values reported in previous studies. These results confirm that integrating deep learning with the Adaptive Whale Optimization Algorithm enhances feature learning and parameter optimization, leading to superior predictive performance and robustness compared with existing approaches.

VI. CONCLUSION AND RECOMMENDATIONS

This study presents an enhanced deep learning model combined with the Adaptive Whale Optimization Algorithm (EDL-AWOA) for psychiatric disorder classification using Reddit mental health text data. A comprehensive Natural Language Processing (NLP) data preprocessing followed by term frequency-inverse document frequency (TF-IDF)-based feature extraction, a two-hidden-layer deep learning structure, and automatically adaptive hyperparameters after multiple iterations were integrated to effectively predict six mental health categories, including Anxiety/Stress, Bipolar, Major Depressive with or without psychotic features, Normal, Personality Disorder, and Substance Abuse or not. The experimental results indicated that the adaptive optimization strategy can optimize the learning procedure and provide better classification accuracy, faster convergence time, and better generalisation performance.

In the set of analysed alpha schedules, the AWOA configuration (0.9 $\leftrightarrow$ 0.1 adaptive) outperformed all others, with an accuracy value of 93.99%, a precision value of 94.01%, and a recall value of 93.99% for the micro and weighted average F1-score values, whilst obtaining results equal in Macro AUC to nearly one at 0.9993. Confusion matrix analysis indicated high classification consistency across all classes, with only a small number of instances in which clinically related disorders were misclassified. Additionally, the ROC analysis confirmed a high discriminative ability (all AUCs were found to need development) for all classes—ranging from 0.995 to 0.997.

Experimental results revealed that EDL-AWOA drastically improved both the DL model and the DL-WOA model in terms of overall performance. The proposed model outperforms DL and DL-WOA by 21.08% and 12.20% (accuracy), respectively, whilst achieving higher scores in precision, recall, F1-score, and Macro AUC in the test dataset. Moreover, the comparison between recent state-of-the-art studies in the field showed that EDL-AWOA obtained the best overall accuracy and AUC values, indicating that adaptive whale-based optimization is a practical approach to effectively solving these mental health classification problems.

The findings confirm that combining adaptive metaheuristic optimization with deep learning yields a robust, adaptable framework with potential for automated psychiatric disorder detection. Thus, the proposed system will make this process more efficient by supporting real-time decision-making for mental health care workers, since they can relatively quickly detect psychopathologies from large volumes of online text data. Nonetheless, future work may explore multilingual datasets, transformer-based embeddings, explainable artificial intelligence methods, and multimodal data sources (speech, clinical records, and behavioral signals) to optimize diagnostic performance and potential utility in the clinic.

An important research direction for the future is to extend the model output from the proposed work on multilingual mental health text classification. Note: This research only applies to English-language Reddit posts, so it does not cover mental health literature in all languages. Future studies should incorporate Arabic and more languages to allow for translation of the model in different ethical and cultural situations since somatic accounts for communicating psychological distress are typically influenced by lexicon, culture, dialect, and social class.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Amman Arab University and the University of Québec for providing the academic environment and research support necessary to conduct this study. The authors also acknowledge the developers and maintainers of the publicly available Reddit mental health datasets and the open-source software libraries used in this research, including Python, PyTorch, Scikit-learn, NLTK, Pandas, and NumPy. Finally, the authors thank the reviewers and editors for their valuable comments and constructive suggestions, which have significantly improved the manuscript's quality, clarity, and rigor.

REFERENCES

- [1] Stein, D. J., Shoptaw, S. J., Vigo, D. V., Lund, C., Cuijpers, P., Bantjes, J., Sartorius, N., & Maj, M. (2022). Psychiatric diagnosis and treatment in the 21st century: Paradigm shifts versus incremental integration. *World Psychiatry*, 21(3), 393-414. <https://doi.org/10.1002/wps.20998>
- [2] O'Dea, B., Batterham, P. J., Braund, T. A., Chakouch, C., Larsen, M. E., Berk, M., Torok, M., Christensen, H., & Glozier, N. (2025). A randomised cross-over trial examining the linguistic markers of depression and anxiety in symptomatic adults. *NPJ Mental Health Research*, 4(1), 30. <https://doi.org/10.1038/s44184-025-00111-8>
- [3] Sun, J., Dong, Q.-X., Wang, S.-W., Zheng, Y.-B., Liu, X.-X., & Lu, L. (2023). Artificial intelligence in psychiatry research, diagnosis, and therapy. *Asian Journal of Psychiatry*, 87, 103705. <https://doi.org/10.1016/j.ajp.2023.103705>
- [4] Nadimi-Shahraki, M. H., Taghian, S., Mirjalili, S., & Faris, H. (2023). A systematic review of the Whale Optimization Algorithm: Theoretical foundation, improvements, and hybridizations. *Archives of Computational Methods in Engineering*, 30, 4113-4159. <https://doi.org/10.1007/s11831-022-09859-7>
- [5] Arasteh, S. T., Han, T., Lotfinia, M., Kuhl, C., Kather, J. N., Truhn, D., & Nebelung, S. (2024). Large language models streamline automated machine learning for clinical studies. *Nature Communications*, 15(1), 1603. <https://doi.org/10.1038/s41467-024-45943-0>
- [6] Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2021). Social media big data analysis for mental health research. In *Big Data Analytics for Healthcare* (pp. 109-143). Elsevier.

- [7] Liu, D., Feng, X. L., Ahmed, F., Shahid, M., & Guo, J. (2022). Detecting and measuring depression on social media using a machine learning approach: Systematic review. *JMIR Mental Health*, 9(3), e27244. <https://doi.org/10.2196/27244>
- [8] Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: Literature review. *Interactive Journal of Medical Research*, 13, e55067. <https://doi.org/10.2196/55067>
- [9] Hameed, S., Nauman, M., Akhtar, N., Fayyaz, M. B., & Nawaz, R. (2025). Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. *Frontiers in Artificial Intelligence*, 8, 1627078. <https://doi.org/10.3389/frai.2025.1627078>
- [10] Senanayake, A., & Liang, Z. (2025). Anxiety detection in Reddit posts through emotion dynamics analysis. In *Proceedings of SCITEPRESS* (pp. 920-927).
- [11] Baydili, İ., Tasci, B., & Tasci, G. (2025). Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. *Behavioral Sciences*, 15(3), 352. <https://doi.org/10.3390/bs15030352>
- [12] Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. R. (2024). Depression symptoms modelling from social media text: An LLM-driven semi-supervised learning approach. *Language Resources and Evaluation*, 58(3), 1013-1041. <https://doi.org/10.1007/s10579-023-09683-4>
- [13] Arif, M., Ameer, I., Bölücü, N., Sidorov, G., Gelbukh, A., & Elangovan, V. (2024). Mental illness classification on social media texts using deep learning and transfer learning. *Computación y Sistemas*, 28(2). DOI not available.
- [14] Nayak, S., Mahapatra, D., Chatterjee, R., Parida, S., & Dash, S. R. (2022). A machine learning approach to analyze mental health from Reddit posts. In *Smart Innovation, Systems and Technologies* (pp. 357-366). DOI not available.
- [15] Hosseinali, F. (2024). Accelerated gradient descent using improved selective backpropagation. *Expert Systems with Applications*, 255, 124426. <https://doi.org/10.1016/j.eswa.2024.124426>
- [16] Moosavi, S. K. R., Akhter, M. N., Zafar, M. H., & Mansoor, M. (2024). Constraint optimization: Solving engineering design problems using Whale Optimization Algorithm (WOA). In S. Mirjalili (Ed.), *Handbook of Whale Optimization Algorithm* (pp. 193-216). Academic Press. <https://doi.org/10.1016/B978-0-32-395365-8.00021-X>
- [17] Scherbakov, D. A., Talwar, A., Desai, P., Roche, K. L., Kwan, B. M., & Iturralde, E. (2025). Natural language processing and social determinants of health in mental health research: Scoping review. *JMIR Mental Health*, 12, e67192. <https://doi.org/10.2196/67192>
- [18] Qi, Q., Hessen, D., Deoskar, T., & Van der Heijden, P. G. M. (2024). A comparison of latent semantic analysis and correspondence analysis of document-term matrices. *Natural Language Engineering*, 30(4), 722-752. <https://doi.org/10.1017/S1351324923000289>
- [19] Sajid, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2024). Neuro-fuzzy random vector functional link neural network for classification and regression problems. *IEEE Transactions on Fuzzy Systems*, 32(5), 2738-2749. <https://doi.org/10.1109/TFUZZ.2023.3330012>
- [20] Rashedi, K. A., Nejatian, S., Parvin, H., Rezaie, V., & Minaei-Bidgoli, B. (2024). Multi-layer perceptron-based classification with missing data. *Journal of Risk and Financial Management*, 17(2), 69. <https://doi.org/10.3390/jrfm17020069>
- [21] Reyad, M., Sarhan, A. M., & Arafa, M. (2023). An enhanced Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, 35(23), 17095-17112.
- [22] Reddy, K., & Saha, A. K. (2022). An enhanced Whale Optimization Algorithm for exploitation capability and stability enhancement. *Heliyon*, 8(10), e11019. <https://doi.org/10.1016/j.heliyon.2022.e11019>
- [23] Utsa, A. D. (2026). Early linguistic pattern of anxiety from social media using interpretable linguistic features: A multi-faceted validation study with author-disjoint evaluation (arXiv:2601.11758). arXiv. <https://arxiv.org/abs/2601.11758>
- [24] Jena, M., Khan, N., Lee, M. Y., & Rho, S. (2026). Neuro-symbolic graph learning for causal inference and continual learning in mental-health risk assessment. *Computer Modeling in Engineering & Sciences*. DOI <https://doi.org/10.32604/cmescs.2025.075119>.
- [25] Kim, S., Imieye, O., & Yin, Y. (2025). Interpretable depression detection from social media text using LLM-derived embeddings (arXiv:2506.06616). arXiv. <https://arxiv.org/abs/2506.06616>