JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

An Interpretable Neuro-Fuzzy Framework for Mental Health Indicator Classification with Contextual Representation and Metaheuristic Optimization

Haeni Budiati [1,*], Sunneng Sandino Berutu [2], Agustinus Rudatyo Himamunanto [1], Faisal Yahya [1], and Mosiur Rahaman [3]

[1] Department of Informatics, Universitas Kristen Immanuel, Sleman 55571, Yogyakarta, Indonesia; e-mail: heni@ukrimuniversity.ac.id; rudatyo@ukrimuniversity.ac.id; faisal.yahya@mail.ukrimuniversity.ac.id
[2] Master's Program in Information Technology, Universitas Duta Bangsa Surakarta, Surakarta 57135, Indonesia; e-mail: sunneng_sandino@udb.ac.id
[3] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand; e-mail: mosiurahaman@gmail.com
* Corresponding Author: Haeni Budiati

Abstract: Social media text provides valuable information for analyzing mental health-related expressions. However, contextual language variation, class imbalance, and limited model interpretability remain important challenges in text classification. This study evaluates different text representations, class-imbalance handling strategies, and metaheuristic optimization within an Adaptive Neuro-Fuzzy Inference System (ANFIS) framework. A total of 19,429 Indonesian-language posts were retained after duplicate removal and label-conflict adjudication. The dataset was partitioned using a fixed group-aware split into 13,600 training, 2,914 validation, and 2,915 test records, covering four categories: Distress, Crisis, Protective, and Situational. The experiments examined lexical-statistical, distributed-semantic, and contextual representations using Plain ANFIS, class-balanced focal ANFIS (CBF-ANFIS), and its metaheuristic-optimized variants based on the Whale Optimization Algorithm (WOA) and Crow Search Algorithm (CSA). Model performance was evaluated across three random seeds and compared with a TF-IDF-based Logistic Regression baseline and direct IndoBERT fine-tuning. Direct IndoBERT achieved the highest test Macro-F1 of 0.7685 ± 0.0046, followed by preliminary Plain ANFIS (0.6942 ± 0.0011) and final CBF-ANFIS (0.6897 ± 0.0054). Although WOA and CSA improved validation Macro-F1, their mean test Macro-F1 scores were slightly lower than that of CBF-ANFIS (0.6890 and 0.6859 versus 0.6897, respectively), indicating no consistent generalization benefit. Overall, direct IndoBERT provided superior classification performance, whereas the neuro-fuzzy framework offered a structured mechanism for tracing rule activation and inference in the latent feature space. Nevertheless, these rules do not provide direct linguistic or clinical interpretations.

Keywords: ANFIS; IndoBERT; Interpretability; Metaheuristic Optimization; Mental Health Text Classification; Neuro-Fuzzy Systems; Social Media; Text Classification.

Received: August, 27th 2026
Revised: October, 20th 2026
Accepted: October, 29th 2026
Published: November, 10th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

The development of Natural Language Processing (NLP) has supported the analysis of mental-health-related indicators from text on digital platforms and social media. User-generated text contains linguistic patterns, emotional expressions, and contextual information that can be used to identify indicators related to psychological conditions. NLP has been applied to different mental-health detection tasks [1], including the identification of depression from social-media data [2] and the detection of depression and suicide-related indicators using deep learning methods [3]. Automated approaches have also been investigated for depression diagnosis from textual information [4]. These studies show the potential of NLP for mental-health-related text analysis. However, differences in language, context, and classification settings remain important factors that can affect the performance of the classification model. In

addition to predictive performance, the application of Artificial Intelligence (AI) in the domain of healthcare also requires transparency regarding the basis of model decision-making [5], [6].

Text representation determines the type of linguistic information available to the classification model. Lexical-statistical representations capture word occurrence and importance, while distributed representations capture semantic relationships between words. In mental-health text classification, linguistic and statistical features have been used for depression-related post detection [7], while word embeddings combined with emotion-related information have been applied to early depression detection [8]. More recent studies have used contextual representations to capture information that depends on the surrounding text. Transformer-based representations combined with metadata and linguistic features have been investigated for depression classification [9], while contextual information combined with mental-state knowledge has been used for early stress and depression detection [10]. These studies indicate that different text representations provide different types of information to the classification model and may therefore affect the resulting classification performance.

Although they offer strong contextual representations, transformer and deep learning models generally have complex decision-making mechanisms that make them difficult to trace [6], [11]. In healthcare applications, high predictive performance alone does not necessarily ensure that model decisions can be understood and trusted [5], [12]. Studies on explainable artificial intelligence (XAI) show that the need for transparency and interpretability is increasing along with the growing use of increasingly complex machine learning models [5], [6]. Moreover, a gap remains between explainable and interpretable deep learning in healthcare NLP because many approaches still rely on post-hoc explanation mechanisms [6]. This condition indicates the need for an approach capable of maintaining data representation capabilities while also providing an interpretable reasoning mechanism [5], [6].

The Adaptive Neuro-Fuzzy Inference System (ANFIS) combines adaptive learning with fuzzy inference through membership functions and fuzzy IF–THEN rules. This structure allows the reasoning process to be traced through membership values and rule activation [13], [14]. In this study, ANFIS operates on reduced numerical text representations. Therefore, interpretability is limited to structured and traceable reasoning in the latent feature space rather than direct linguistic or psychological interpretation of the original text.

The performance of ANFIS depends on the configuration and optimization of its model parameters. Mehrabi et al. [15] demonstrated the optimization of ANFIS using several metaheuristic techniques, while Ahmadianfar et al. [16] employed conjoined metaheuristic algorithms to improve ANFIS parameter optimization. More recently, Wang et al. [17] systematically evaluated hybrid ANFIS frameworks using evolutionary and swarm-based optimization algorithms. These studies show that metaheuristic methods provide alternative mechanisms for exploring the ANFIS parameter space, although their effectiveness depends on the optimization problem and model configuration.

Metaheuristic optimization is one approach that can be used to overcome the limitations of parameter search. Hybridization of ANFIS with evolutionary algorithms and swarm-based metaheuristics enables broader exploration of the parameter space [15]–[17]. In the context of this study, the Whale Optimization Algorithm (WOA) and Crow Search Algorithm (CSA) are considered because they have different exploration–exploitation characteristics [18], [19]. However, metaheuristic optimization cannot be assumed to always improve ANFIS performance. WOA has been reported to have several limitations, including a tendency toward search bias and performance degradation under certain conditions [18], while various CSA developments have been directed toward improving search capabilities, reducing premature convergence, and improving the exploration–exploitation balances [19], [20]. Thus, the effectiveness of optimization needs to be evaluated empirically against the characteristics of the representation and search space used.

Based on this literature, there is a methodological gap that has not yet been fully addressed. Mental-health NLP research has more often utilized contextual representations to improve predictive performance [1], [4], [9], [10], while the interpretability of decisions remains relatively limited [5], [6], [11], [12]. On the other hand, ANFIS-metaheuristic research generally treats optimization as a mechanism for improving prediction quality [15]–[17], but the interaction between the level of semantic richness of text representations and the optimization process on the structure of fuzzy reasoning has rarely been evaluated in an integrated manner. XAI approaches in healthcare and mental-health NLP also still largely use post-hoc

explanations for complex models [5], [6], [11], [12], while fuzzy logic provides an intrinsic interpretability mechanism through membership functions and linguistic rules [13], [14]. Therefore, there remains a need for a framework that systematically integrates text representation, neuro-fuzzy reasoning, metaheuristic optimization, predictive performance, and intrinsic interpretability.

To address this gap, this study proposes an interpretable neuro-fuzzy framework for text-based mental health indicator classification. TF-IDF, Word2Vec, and IndoBERT are used to represent three levels of information, namely lexical-statistical, distributed-semantic, and contextual representation, which are subsequently integrated with ANFIS. WOA and CSA are used to evaluate the effects of optimization strategies with different exploration–exploitation characteristics on the ANFIS learning process. Unlike studies that only compare representation accuracy or assume that metaheuristic optimization will improve performance, this study evaluates the representation–optimization–performance–interpretability relationship in an integrated manner. Interpretability is analyzed intrinsically through membership functions and fuzzy IF–THEN rules to identify how representation and optimization strategies affect not only predictive performance but also the model's decision-making mechanism.

The main contributions of this study are: (1) developing an integrated framework that connects text representation, neuro-fuzzy reasoning, metaheuristic optimization, and interpretability; (2) evaluating the effects of the semantic richness levels of TF-IDF, Word2Vec, and IndoBERT on ANFIS learning; (3) analyzing the effects of WOA and CSA on the quality of ANFIS learning without assuming that optimization always produces improvements over the ANFIS baseline; and (4) evaluating latent-space interpretability through membership values and fuzzy-rule activation. Accordingly, this study is directed toward obtaining a more comprehensive understanding of the balance between contextual understanding, predictive performance, optimization effectiveness, and interpretable reasoning.

2. Related Work

2.1. NLP and Text Representation for Mental Health

The increasing availability of user-generated text on social media has supported the development of NLP methods for identifying mental-health-related indicators. NLP has been applied to different mental-health detection tasks [1], including depression detection from social-media data [2] and depression and suicide-related detection using deep learning methods [3]. Automated approaches have also been investigated for depression diagnosis [4]. These studies show that social-media text provides relevant information for mental-health-related classification, although the expressions being analyzed may vary across tasks and contexts.

Mental-health-related text has been represented using different types of linguistic information. Linguistic and statistical features have been used for depression-related post detection [7], while word embeddings combined with emotion-related information have been applied to early depression detection [8]. More recent approaches have used transformer-based representations to capture contextual information. Transformer-based representations combined with auxiliary features have been investigated for depression classification [9], while contextual information combined with mental-state knowledge has been used for early stress and depression detection [10]. These studies indicate that lexical, distributed-semantic, and contextual representations provide different types of information for mental-health-related text classification.

Recent studies have investigated the effects of preprocessing and contextual representation on Indonesian text classification. Budianoor et al. [21] evaluated different preprocessing strategies for IndoBERT fine-tuning on informal Indonesian text and showed that preprocessing choices affected classification performance. Aqiilah et al. [22] investigated IndoBERTweet for sentiment classification of Indonesian application reviews and demonstrated the effectiveness of contextual representation for informal text. Darnoto and Firmawan [23] examined language-similarity-guided transfer fine-tuning across Indonesian regional languages, showing the relevance of language adaptation in transformer-based classification.

These studies indicate that preprocessing, contextual representation, and transfer learning can affect text classification performance. However, they mainly focused on sentiment classification and did not examine the integration of contextual representation, neuro-fuzzy

classification, class-imbalance handling, and metaheuristic optimization for mental-health indicator classification. Therefore, the present study evaluates these components within a consistent experimental framework to examine their effects on classification performance and latent-space fuzzy-rule interpretability.

2.2. Class Imbalance and Interpretability in Mental-Health Classification

Mental-health-related text classification also presents challenges related to class distribution and model transparency. An unequal class distribution can affect the learning process because less represented categories contribute fewer observations during model training. In addition, increasing model complexity makes the classification process more difficult to trace. This issue is particularly relevant in healthcare-related applications, where predictive performance needs to be considered together with the transparency of the decision process [5], [6], [12].

Explainable AI has been applied to provide additional information about the decisions produced by complex predictive models. Huang et al. [6] distinguished explainable and interpretable approaches in healthcare NLP. XAI methods have also been applied to identify linguistic characteristics associated with mental-health conditions [11]. However, many complex models rely on post-hoc explanations generated after the prediction has been produced. As a result, the explanation mechanism remains separate from the inference process.

Fuzzy systems provide a different mechanism through membership functions and rule activation [13], [14]. This structure allows the inference process to be traced from the input features to the activated fuzzy rules and model output. However, when the inputs are latent numerical representations, the resulting fuzzy rules cannot be directly interpreted as linguistic or psychological concepts. Therefore, interpretability in this study refers to structured and traceable fuzzy-rule activation in the latent feature space rather than direct linguistic interpretation of the original text.

Class imbalance presents an additional problem because improvement in overall performance does not necessarily indicate similar improvements across all classes. Therefore, the effect of imbalance handling needs to be evaluated separately before performance changes are attributed to the representation, neuro-fuzzy model, or optimization algorithm. In this study, this separation provides a controlled basis for examining the contribution of class-sensitive learning to the ANFIS classification process.

2.3. Neuro-Fuzzy Modeling and Metaheuristic Optimization

ANFIS combines adaptive learning with fuzzy inference through membership functions and IF–THEN rule structures. This structure enables model parameters to be learned from data while maintaining a rule-based inference process. The premise parameters determine the membership functions, while the consequent parameters determine the output associated with the activated rules. This structure provides a basis for examining how different text representations are processed through structured fuzzy reasoning.

The optimization of ANFIS parameters becomes more complex as the dimensionality of the input and the number of model parameters increase. Previous studies have therefore investigated metaheuristic algorithms as alternative mechanisms for ANFIS parameter optimization. Mehrabi et al. [15] evaluated several metaheuristic techniques for optimizing ANFIS parameters, while Ahmadianfar et al. [16] investigated conjoined metaheuristic algorithms for improving the ANFIS optimization process. Wang et al. [17] further evaluated hybrid ANFIS frameworks using evolutionary and swarm-based optimization algorithms. These studies show that metaheuristic algorithms can provide different search mechanisms for exploring the ANFIS parameter space.

Among population-based optimization methods, WOA performs parameter search through exploration and exploitation mechanisms [18], [24]. However, its optimization behavior can be affected by search bias and the characteristics of the search problem [18]. CSA follows a different search mechanism based on the memory and awareness behavior of search agents [19]. Subsequent developments of CSA have focused on improving its search capability and reducing limitations such as premature convergence [20], [25]. Therefore, WOA and CSA provide different optimization mechanisms for evaluating the ANFIS parameter search process.

The different search characteristics of WOA and CSA provide the basis for their evaluation in this study. Both algorithms are applied under the same ANFIS configuration and optimization setting to determine whether their search mechanisms improve validation-based parameter optimization. However, an improvement in validation performance does not necessarily indicate better predictive performance on unseen data. Therefore, the optimized models are subsequently evaluated on the held-out test set to determine whether the validation improvements are maintained during generalization.

2.4. Research Gap and Study Positioning

Previous studies have addressed different components of mental-health text classification from different methodological perspectives. Mental-health NLP studies have investigated linguistic, semantic, and contextual information for depression, stress, and related classification tasks [1]–[4], [7]–[10]. Studies on explainable AI have emphasized the importance of understanding the reasoning behind predictions produced by complex models [5], [6], [11], [12]. Fuzzy and neuro-fuzzy approaches provide structured inference through membership functions and rule-based reasoning [13], [14], while metaheuristic algorithms have been investigated as alternative mechanisms for ANFIS parameter optimization [15]–[17]. However, these components have generally been investigated separately rather than evaluated within the same experimental framework.

First, different text representations provide different types of information to the classification model. Linguistic and statistical features have been used for depression-related post detection [7], word embeddings have been combined with emotional information for early depression detection [8], and transformer-based representations have been investigated for contextual mental-health classification [9], [10]. However, the effects of lexical-statistical, distributed-semantic, and contextual representations on the same neuro-fuzzy classifier remain less systematically examined. Second, class-imbalance handling can affect classification performance, particularly for less represented classes, and its contribution therefore needs to be evaluated separately from changes produced by text representation and model optimization. Third, fuzzy inference provides a structured reasoning mechanism through membership functions and rule activation [13], [14]. However, when latent numerical representations are used as inputs, the resulting rule activation reflects reasoning in the latent feature space rather than direct linguistic or psychological explanations.

Metaheuristic optimization introduces another methodological consideration. Previous studies have investigated different metaheuristic strategies for ANFIS parameter optimization [15]–[17]. WOA performs population-based search through exploration and exploitation mechanisms [18], whereas CSA uses a memory-guided search mechanism [19]. Subsequent developments of these algorithms have focused on improving their search and convergence characteristics [20]–[26]. However, improvements obtained during validation-based optimization do not necessarily indicate better generalization. Therefore, optimization performance needs to be verified independently using unseen test data.

To address these gaps, this study evaluates text representation, class-imbalance handling, and metaheuristic optimization within a controlled ANFIS-based experimental framework. TF-IDF, Word2Vec, and IndoBERT represent lexical-statistical, distributed-semantic, and contextual information, respectively, and are first compared under a common ANFIS setting. Class-imbalance handling is then examined separately, followed by the application of WOA and CSA to ANFIS parameter optimization. The resulting models are evaluated on a held-out test set to determine whether validation-based improvements translate into better generalization. Conventional classifiers and direct IndoBERT fine-tuning are also included as predictive reference baselines using the same fixed data split. Beyond predictive performance, the evaluation considers the extent to which ANFIS provides structured and traceable inference through rule activation in the latent feature space, without treating such inference as direct linguistic or psychological explanation.

3. Method

3.1. Study Design and Experimental Control

This study employed a multiclass text-classification framework to evaluate text representations, direct IndoBERT fine-tuning, and ANFIS-based models incorporating class-

imbalance handling and metaheuristic optimization. All experiments used fixed data partitions comprising 13,600 training, 2,914 validation, and 2,915 test records. The training set was used for model development, the validation set for model selection and hyperparameter tuning, and the test set exclusively for final performance evaluation.

The experimental procedure consisted of preliminary screening and final evaluation. During preliminary screening, TF-IDF, Word2Vec, IndoBERT Mean, and IndoBERT CLS representations were compared using Plain ANFIS. The selected IndoBERT Mean representation was subsequently used to examine different loss functions, including cross-entropy, class-balanced focal loss, label smoothing, and logit adjustment. CBF-ANFIS achieved the highest mean validation Macro-F1 and was selected for further evaluation. In the final phase, CBF-ANFIS was retrained using three random seeds (42, 52, and 62). Because the preliminary and final experiments did not use identical training configurations, their results were treated as separate experimental stages rather than a controlled loss-function ablation.

The final CBF-ANFIS checkpoints served as common references for evaluating WOA- and CSA-optimized variants. Both optimizers independently adjusted the antecedent centers and log-sigma parameters while keeping the consequent parameters fixed. Model selection was based on validation Macro-F1, whereas the selected models were evaluated on the held-out test set to assess generalization. Results were reported as mean ± standard deviation across the three seeds, except for applicable single-run baselines.

3.2. Dataset Preparation and Annotation

3.2.1. Data Collection

Indonesian-language posts were collected from X (formerly Twitter) between September 5, 2010, and April 18, 2026, using a third-party crawling agent. Four keyword groups were used for retrieval: Protective (*senang*, *bahagia*, *semangat*, *pulih*, *optimis*, and *bersyukur*), Situational (*tugas*, *lembur*, *skripsi*, deadline, and *kerja*), Distress (*capek*, *kecewa*, *takut*, burnout, overthinking, and stress), and Crisis (*menyerah*, *tidak berharga*, *bunuh diri*, *putus asa*, and *ingin menghilang*). These keywords served exclusively as retrieval criteria and were not used for automatic label assignment. Posts were included if they were written in Indonesian and matched at least one retrieval keyword; retweets were excluded. No specific spam-detection procedure was applied.

3.2.2. Text Preprocessing

The collected posts underwent text cleaning and normalization before feature extraction. URLs and hyperlinks were removed, and user mentions were removed or replaced during cleaning. The text was converted to lowercase, punctuation and special characters were removed, and informal or non-standard expressions were normalized. Stopword removal and stemming were subsequently applied to reduce irrelevant terms and morphological variations. The resulting text was used for TF-IDF, Word2Vec, and IndoBERT representation, as described in Section 3.3.

3.2.3. Class Definitions and Annotation

The dataset was annotated using four predefined categories. Distress covered psychological or emotional strain, including stress, anxiety, sadness, exhaustion, worry, loneliness, and loss of motivation. Crisis represented more severe psychological risk, including suicidal expressions, self-harm, extreme hopelessness, and severe loss of control. The distinction between these two categories was based primarily on the severity and risk expressed in each post. Protective represented coping and recovery resources, such as optimism, self-care, help-seeking, social support, and spiritual coping. Situational covered external stressors, including academic workload, deadlines, financial problems, employment, relationships, and family demands. Labels were assigned based on the overall context of each post rather than individual retrieval keywords.

Three undergraduate students independently annotated each post using these predefined categories. Two annotators were eighth-semester Psychology students from the same university, while the third was a sixth-semester Informatics student from a different university. None was considered a clinical or mental-health professional. Each complete post was assessed because it could contain indicators associated with multiple categories. Inter-annotator agreement was measured using Fleiss' Kappa [27], yielding a score of 0.85.

Of the 19,610 annotated posts, 16,402 (83.64%) received unanimous agreement and 3,034 (15.47%) received majority agreement. The remaining 174 posts (0.89%), for which all

three annotators assigned different labels, underwent consensus adjudication involving the three annotators and three lecturers. All 174 posts received final labels and were retained.

3.2.4. Data Quality Audit and Final Dataset

Following adjudication, a separate data-quality audit examined identical texts with inconsistent labels. The audit identified 45 groups requiring manual review. Twelve groups were retained after 15 individual label corrections, while 33 groups comprising 181 records were excluded. These exclusions involved 31 groups containing out-of-scope content, one unrelated group, and one group that could not be reliably assigned to a single category. The resulting dataset contained 19,429 posts, comprising 5,452 Distress, 4,232 Crisis, 3,668 Protective, and 6,077 Situational records. A group-aware split with seed 42 partitioned the dataset into 13,600 training, 2,914 validation, and 2,915 test records, as summarized in Table 1. These fixed partitions were maintained across all experiments.

Table 1. Dataset Distribution across Data Partitions

Partition	Distress	Crisis	Protective	Situational	Total
Training	3,816	2,962	2,568	4,254	13,600
Validation	818	634	550	912	2,914
Test	818	636	550	911	2,915
Total	5,452	4,232	3,668	6,077	19,429

3.3. Text Representation

This study evaluated TF-IDF, Word2Vec, and IndoBERT to represent lexical, distributed semantic, and contextual information, respectively. Four representations were generated: TF-IDF, Word2Vec, IndoBERT Mean, and IndoBERT CLS. Since their original dimensions differed, each representation was reduced to 64 dimensions before ANFIS classification. All dimensionality-reduction and scaling parameters were fitted on the training set and then applied unchanged to the validation and test sets.

3.3.1. TF-IDF

The TF-IDF representation combined word and character features. Word n-grams of 1–2 were limited to 15,000 features, while character n-grams of 3–5 were limited to 10,000 features, resulting in a maximum of 25,000 features. Truncated SVD reduced the sparse representation to 64 dimensions, followed by feature scaling.

3.3.2. Word2Vec

Word2Vec was trained on the training corpus using Skip-Gram with 300-dimensional embeddings, a context window of 5, 15 negative samples, 40 epochs, and min_count = 1. Document vectors were generated using TF-IDF-weighted aggregation of token embeddings. PCA then reduced the vectors to 64 dimensions, followed by feature scaling.

3.3.3. IndoBERT

IndoBERT representations were generated using indobenchmark/indobert-large-p2 with a maximum sequence length of 128 tokens and a batch size of 8. Mean pooling and [CLS] representations were extracted from the same forward pass. Each representation was independently reduced to 64 dimensions using PCA fitted on the training set and was then scaled. No task-adaptive pretraining was performed.

3.3.4. Framework Integration

The four 64-dimensional representations were used separately as inputs to ANFIS to evaluate their effect on classification performance. The membership functions transformed the input features into fuzzy membership values, while the fuzzy rules combined these values to produce four-class predictions. Plain ANFIS was first evaluated across the four representations. The selected representation was then used to evaluate class-imbalance handling using class-balanced focal loss.

WOA and CSA were subsequently evaluated to optimize the ANFIS antecedent parameters. Both methods optimized the membership-function centers and log-sigmas, while the consequent parameters remained fixed. Validation Macro-F1 was used as the fitness function, and the test set was not involved during optimization.

Therefore, the classification pipeline consisted of text preprocessing, text representation, dimensionality reduction to 64 features, and ANFIS-based classification using fuzzy membership functions and rules. In this process, text representation encodes the input text, dimensionality reduction provides a common feature space, ANFIS performs the classification, and WOA or CSA optimizes the antecedent parameters.

3.4. Baseline Models

Logistic Regression (LR) was used as a conventional baseline. The classifiers were evaluated using fixed 64-dimensional TF-IDF, Word2Vec, IndoBERT Mean, and IndoBERT CLS representations. The regularization parameter was selected from the candidate set $C \in \{0.01,0.1,1,10\}$ based on validation performance. All experiments were conducted using three random seeds: 42, 52, and 62.

Direct IndoBERT fine-tuning was evaluated separately to compare end-to-end transformer learning with classifiers trained on fixed text representations. The model used (indobenchmark/indobert-large-p2), with normalized text as input. The maximum sequence length was set to 128 tokens, and the per-device batch size was 8. Gradient accumulation was set to 2, resulting in an effective batch size of 16. The model was optimized using AdamW with a learning rate of 2×10^{-5}, weight decay of 0.01, a warm-up ratio of 0.10, and a maximum gradient norm of 1.0. Mixed-precision training was also applied [28].

Training was limited to 10 epochs, with an early-stopping patience of 3 epochs. For each random seed (42, 52, and 62), the checkpoint achieving the highest validation Macro-F1 score was selected. The test set was evaluated only after checkpoint selection to avoid using test performance for model selection.

3.5. ANFIS Classifier

The neuro-fuzzy classifier used an Adaptive Neuro-Fuzzy Inference System (ANFIS) with 12 fuzzy rules, a 64-dimensional input, and four output classes. For an input vector $(x \in R^{64})$, the antecedent of rule r was characterized by a center c_{rj} and a positive scale σ_{rj}. The Gaussian membership value of feature x_j for rule (r) was defined as in Eq.(1).

$$\mu_r(z_j) = exp\left(-\frac{(z_j - c_{(rj)})^2}{(2\sigma^2_{(rj)})}\right) \quad (1)$$

where the positive scale was parameterized using the trainable log-scale in Eq.(2),

$$\sigma_{rj} = \exp(s_{rj}) \quad (2)$$

where s_{rj} denotes the trainable log-sigma parameter. This parameterization ensures that the Gaussian scale σ_{rj} remains positive during model training.

To improve numerical stability across the 64 input dimensions, rule activation was computed in the log domain and normalized across the 12 rules. The normalized firing strength of rule (r) was therefore expressed as in Eq.(3).

$$\overline{w_r}(\mathbf{x}) = \frac{\exp\left(\frac{1}{64}\sum_{j=1}^{64}\log\mu_{rj}(x_j)\right)}{\sum_{k=1}^{12}\exp\left(\frac{1}{64}\sum_{j=1}^{64}\log\mu_{kj}(x_j)\right)} \quad (3)$$

As shown in Eqs. (1)–(3), the antecedent layer measures the relationship between each input feature and the rule centers and converts these membership values into normalized rule activations.

Each rule used a first-order consequent for each output class. The input vector was first augmented with a constant term as in Eq.(4).

$$\tilde{\mathbf{x}} = [1, x_1, x_2, \dots, x_{64}]^T \quad (4)$$

For rule (r) and class (c), the rule-specific output was computed as in Eq.(5).

$$f_{rc}(\mathrm{x}) = \mathrm{p}_{\mathrm{rc}}^{\mathrm{T}}\tilde{\mathrm{x}} \quad (5)$$

where p_{rc} represents the consequent parameters. The final ANFIS logit for class (c) was obtained by combining the rule-specific outputs according to their normalized firing strengths as in Eq.(6):

$$z_c(\mathrm{x}) = \sum_{r=1}^{12} \overline{w_r}(\mathrm{x}) f_{rc}(\mathrm{x}), \qquad c = 1, \ldots, 4 \tag{6}$$

Thus, Eqs. (4) – (6) describe how the activated fuzzy rules are combined to produce the four class logits. The antecedent centers were initialized using MiniBatchKMeans centroids, while the initial scales were estimated from within-cluster feature variation with a sigma scale of 1.0.

Plain ANFIS was trained using AdamW with a learning rate of 0.0008, weight decay of 0.0001, dropout of 0.1, and batch size of 64. Unweighted cross-entropy was used as the loss function. Training was limited to 100 epochs with an early-stopping patience of 10, and the checkpoint with the highest validation Macro-F1 was retained. Plain ANFIS was evaluated using TF-IDF, Word2Vec, IndoBERT Mean, and IndoBERT CLS representations. Based on validation performance, IndoBERT Mean was subsequently used for the controlled class-imbalance and metaheuristic experiments.

3.6. Imbalance Handling

During preliminary screening, three imbalance-handling strategies were evaluated using the fixed IndoBERT Mean representation: class-balanced focal (CBF) loss, label smoothing, and logit adjustment. Plain ANFIS with unweighted cross-entropy (CE) served as the reference. The strategy for subsequent investigation was selected based on mean validation Macro-F1 [28]–[30]. CBF loss combines class-balanced weighting and focal modulation to account for class frequencies and reduce the contribution of confidently classified samples. The effective-number weighting parameter was set to $\beta = 0.999$, and the focal parameter to $\gamma = 2.0$. For class c containing n_c training samples, its weight was calculated as:

$$\alpha_c = \frac{1-\beta}{1-\beta^{n_c}} \tag{7}$$

The class weights were normalized before training. For sample i, with true class y_i and predicted probability p_{i,y_i} assigned to that class, the CBF loss was defined as:

$$\mathcal{L}_{\mathcal{CBF}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha_{y_i}(1-p_{i,y_i})^{\gamma} \log\left(p_{i,y_i}\right) \tag{8}$$

Following preliminary screening, CBF-ANFIS was retrained using the finalized configuration, with a batch size of 256, a learning rate of 0.001, weight decay of 0.0001, a maximum of 100 epochs, and early-stopping patience of 12. The preliminary and final results were treated as separate experimental stages because their training configurations were not identical.

A controlled ablation study subsequently compared CBF, CE, label smoothing (LS), and logit adjustment (LA) under the same finalized training protocol. The IndoBERT Mean 64D representation, 12-rule ANFIS architecture, data partitions, experimental seeds, and core training hyperparameters were held constant to isolate the effect of the training objective. The preliminary screening, final model evaluation, and controlled ablation results are reported separately in Sections 4.3, 4.4, and 4.8, respectively.

3.7. Metaheuristic Optimization of CBF-ANFIS

WOA and CSA were independently applied to the selected CBF-ANFIS checkpoints to optimize the antecedent parameters while keeping the consequent parameters fixed. With 12 rules and 64 input features, the optimization involved 768 rule-center parameters and 768 log-sigma parameters, resulting in a 1,536-dimensional search vector [32]:

$$\boldsymbol{\theta} = [\mathrm{vec}(C), \mathrm{vec}(\log \Sigma)] \in \mathbb{R}^{1536} \tag{9}$$

where C and $\log \Sigma$ denote the rule centers and log-sigma parameters, respectively.

The search bounds were defined relative to the source CBF-ANFIS checkpoint. Rule centers were allowed to vary within ±0.25 times the standard deviation of each training feature, whereas log-sigma parameters could vary within ±0.35 of their checkpoint values:

$$c_{r,d}^{L} = c_{r,d}^{CBF} - 0.25s_d, \quad c_{r,d}^{U} = c_{r,d}^{CBF} + 0.25s_d$$
$$l_{r,d}^{L} = l_{r,d}^{CBF} - 0.35, \quad l_{r,d}^{U} = l_{r,d}^{CBF} + 0.35 \tag{10}$$

where s_d denotes the standard deviation of training feature d, $l_{r,d} = \log(\sigma_{r,d})$, and the superscripts L and U indicate the lower and upper search bounds, respectively.

The initial population consisted of the source checkpoint as its first candidate and nine additional candidates sampled uniformly within the search bounds:

$$\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{CBF}, \qquad \boldsymbol{\theta}_i \sim U(\boldsymbol{\theta}^L, \boldsymbol{\theta}^U), \quad i = 1, \ldots, 9 \tag{11}$$

Both optimizers used validation Macro-F1 as the fitness function, with no optimizer-level early stopping. The best validation solution from each run was subsequently evaluated on the held-out test set. Each run involved 210 fitness evaluations (10 initial evaluations and 200 evaluations across 20 iterations), yielding 630 evaluations per optimizer and 1,260 evaluations overall.

3.7.1. Whale Optimization Algorithm

WOA used a spiral coefficient of b = 1.0. Its control parameter a decreased linearly from 2 to 0 over 20 iterations [30]:

$$a(t) = 2\left(1 - \frac{t}{T-1}\right) \tag{12}$$

where (T=20) is the maximum number of iterations and ($t = 0, \ldots, T-1$). The spiral updating mechanism was defined as:

$$\boldsymbol{\theta}(t+1) = D' e^{bl} cos(2\pi l) + \boldsymbol{\theta}^*(t) \tag{13}$$

where $\boldsymbol{\theta}^*(t)$ represents the best validation solution, $D' = |\boldsymbol{\theta}^*(t) - \boldsymbol{\theta}(t)|$, and l is randomly sampled from [−1, 1]. Candidate solutions exceeding the search bounds were clipped to the corresponding limits.

3.7.2. Crow Search Algorithm

CSA used an awareness probability of AP=0.10, and a flight length of FL=2.0. When crow j was not aware of being followed, crow i moved toward the remembered position of crow j according to [33]:

$$\boldsymbol{\theta}_i^{t+1} = \boldsymbol{\theta}_i^t + r_i FL\left(\boldsymbol{m}_j^t - \boldsymbol{\theta}_i^t\right) \tag{14}$$

where r_i is a random number in ([0,1]) and $\boldsymbol{m}_j^t$ denotes the remembered position of crow j Otherwise, a random candidate was generated within the search bounds. Each crow's memory was updated when its new candidate achieved a higher validation Macro-F1.

3.8. Evaluation Metrics

Model performance was evaluated using accuracy, balanced accuracy, Macro-F1, Matthews correlation coefficient (MCC), macro-ROC-AUC, and macro-PR-AUC. Balanced accuracy and Macro-F1 were used to account for differences in class distribution, while MCC was used to assess the overall agreement between predicted and true labels [34], [35]. Macro ROC-AUC and macro PR-AUC were used to evaluate class discrimination across different decision thresholds, providing complementary evaluations of classification performance, particularly in the presence of class imbalance [36].

Precision, recall, and F1-score were also reported for each class to examine class-specific performance, particularly for Distress, Crisis, Protective, and Situational [35]. Confusion matrices were used to identify the main misclassification patterns. For each random seed (42, 52, and 62), all evaluation metrics and confusion matrices were computed from the same held-out test predictions. Aggregate performance was reported as mean ± standard deviation across the three seeds, while individual confusion matrices represented seed-specific results.

3.9. Statistical and Error Analysis

Statistical comparisons were performed using the configuration selected from the validation set for each method. No model was reselected based on test performance. Paired bootstrap analysis with 1,000 resamples was applied to the locked test predictions to evaluate pairwise differences in Macro-F1. The resulting 95% confidence intervals were used to determine the stability of the observed differences.

Error analysis was performed using the same locked test predictions. Per-class precision, recall, and F1-score, together with confusion matrices, were examined to identify the main misclassification patterns. Cross-seed prediction agreement was also analyzed for CBF-ANFIS, WOA-ANFIS, and CSA-ANFIS to determine whether differences in aggregate performance were accompanied by changes in individual predictions. Qualitative interpretations of classification errors were made only after examining the corresponding text.

3.10. Experimental Configuration

The main experimental configurations are summarized in Table 2, including data partitioning, text representations, ANFIS architectures, metaheuristic settings, experimental seeds, evaluation metrics, and statistical analysis. All experiments followed the evaluation protocol described in Section 3.1, with the validation set used for model selection and optimization, while the held-out test set was reserved for final performance evaluation.

Table 2. Experimental Configuration

Component	Configuration
Data split	Group-aware split: 13,600 training, 2,914 validation, and 2,915 test samples; split seed = 42
Text representations	TF-IDF, Word2Vec, IndoBERT Mean, and IndoBERT CLS; each reduced to 64 dimensions
Plain ANFIS	12 fuzzy rules; AdamW; learning rate = 0.0008; weight decay = 0.0001; dropout = 0.1; batch size = 64; maximum 100 epochs; patience = 10; unweighted cross-entropy
CBF-ANFIS	IndoBERT Mean; β=0.999; γ=2.0; learning rate = 0.001; weight decay = 0.0001; batch size = 256; maximum 100 epochs; patience = 12
WOA-ANFIS	Population size = 10; 20 iterations; (b=1.0); center radius factor = 0.25; log-sigma radius = 0.35; antecedent parameters optimized; consequent parameters fixed
CSA-ANFIS	Population size = 10; 20 iterations; (AP=0.10); (FL=2.0); center radius factor = 0.25; log-sigma radius = 0.35; antecedent parameters optimized; consequent parameters fixed
Metaheuristic fitness	Validation Macro-F1
Experimental seeds	42, 52, and 62
Evaluation metrics	Accuracy, balanced accuracy, Macro-F1, MCC, macro-ROC-AUC, and macro-PR-AUC
Statistical analysis	Paired bootstrap with 1,000 resamples

4. Results and Discussion

4.1. Model Performance Comparison

Table 3 summarizes the held-out test performance of the reference baselines, preliminary Plain ANFIS, and final-stage ANFIS models. Direct IndoBERT achieved the highest Macro-F1 (0.7685 ± 0.0046), followed by Plain ANFIS (0.6942 ± 0.0011) and CBF-ANFIS (0.6897 ± 0.0054). However, Plain ANFIS was evaluated during preliminary screening under a different training configuration; its results are therefore included for descriptive rather than controlled comparison. Direct IndoBERT and TF-IDF + LR serve as independent predictive reference baselines, whereas WOA-ANFIS and CSA-ANFIS are compared with their seed-matched final CBF-ANFIS checkpoints.

Within the finalized experiments, WOA-ANFIS and CSA-ANFIS achieved Macro-F1 scores of 0.6890 ± 0.0089 and 0.6859 ± 0.0058, respectively, compared with 0.6897 ± 0.0054 for CBF-ANFIS. Despite using the same optimization settings and seed-matched source checkpoints, neither optimizer consistently improved held-out test performance. Overall, direct IndoBERT provided the strongest predictive performance, while ANFIS offered a framework for evaluating fuzzy-rule-based classification and antecedent optimization. The effects of different loss functions under identical finalized training configurations are examined separately in Section 4.8.

Table 3. Held-Out Test Performance by Experimental Stage

Experimental stage	Model	Accuracy	Balanced Acc.	Macro-F1	MCC	ROC-AUC	PR-AUC
Reference baselines	TF-IDF + LR	0.6775	0.6651	0.6645	0.5651	0.8860	0.7533
Reference baselines	Direct IndoBERT	0.7663 ± 0.0041	0.7706 ± 0.0070	0.7685 ± 0.0046	0.6856 ± 0.0065	0.9327 ± 0.0076	0.8489 ± 0.0146
Preliminary screening	Plain ANFIS	0.6938 ± 0.0010	0.6901 ± 0.0007	0.6942 ± 0.0011	0.5855 ± 0.0016	0.8976 ± 0.0056	0.7799 ± 0.0093
Final evaluation	CBF-ANFIS	0.6895 ± 0.0051	0.6874 ± 0.0055	0.6897 ± 0.0054	0.5801 ± 0.0071	—	—
Final evaluation	WOA-ANFIS	0.6883 ± 0.0088	0.6882 ± 0.0087	0.6890 ± 0.0089	0.5788 ± 0.0119	—	—
Final evaluation	CSA-ANFIS	0.6854 ± 0.0058	0.6841 ± 0.0067	0.6859 ± 0.0058	0.5747 ± 0.0082	—	—

Note: Values are reported as mean ± standard deviation across three runs, where applicable.

4.2. Effect of Text Representation

The four 64-dimensional text representations were evaluated using the same Plain ANFIS architecture. As shown in Table 4, IndoBERT Mean achieved the highest validation Macro-F1 (0.7081 ± 0.0058), followed by IndoBERT CLS (0.6938 ± 0.0097), TF-IDF (0.6907 ± 0.0007), and Word2Vec (0.6598 ± 0.0054). Accordingly, IndoBERT Mean was selected for subsequent imbalance-handling and metaheuristic optimization experiments.

The held-out test results showed a consistent ranking, with IndoBERT Mean achieving the highest Macro-F1 (0.6942 ± 0.0011), balanced accuracy (0.6901 ± 0.0007), and MCC (0.5855 ± 0.0016). These findings indicate that contextual representations provided more effective input features for Plain ANFIS than the other representations evaluated under the same configuration.

Table 4. Plain ANFIS Performance across Text Representations

Representation	Validation Macro-F1	Test Accuracy	Test Balanced Acc	Test Macro-F1	Test MCC
TF-IDF	0.6907 ± 0.0007	0.6874 ± 0.0074	0.6740 ± 0.0082	0.6737 ± 0.0094	0.5787 ± 0.0098
Word2Vec	0.6598 ± 0.0054	0.6616 ± 0.0055	0.6498 ± 0.0042	0.6547 ± 0.0051	0.5413 ± 0.0075
IndoBERT Mean	0.7081 ± 0.0058	0.6938 ± 0.0010	0.6901 ± 0.0007	0.6942 ± 0.0011	0.5855 ± 0.0016
IndoBERT CLS	0.6938 ± 0.0097	0.6866 ± 0.0062	0.6824 ± 0.0073	0.6857 ± 0.0071	0.5759 ± 0.0087

Note: Values are reported as mean ± standard deviation across three runs, where applicable.

4.3. Preliminary Screening of Imbalance-Handling Strategies

The preliminary screening compared four training objectives using the IndoBERT Mean representation: unweighted cross-entropy, class-balanced focal loss, label smoothing, and logit adjustment. As shown in Table 5, CBF-ANFIS achieved the highest validation Macro-F1 (0.7108 ± 0.0056), followed by label smoothing (0.7101 ± 0.0076), logit adjustment (0.7093 ± 0.0069), and Plain ANFIS (0.7081 ± 0.0058). Accordingly, CBF-ANFIS was selected for subsequent investigation.

On the held-out test set, label smoothing achieved the highest Macro-F1 (0.6958 ± 0.0058), closely followed by CBF-ANFIS (0.6954 ± 0.0045), Plain ANFIS (0.6942 ± 0.0011), and logit adjustment (0.6924 ± 0.0061). The relatively small differences indicate that the preliminary screening did not establish a substantial predictive advantage for any individual imbalance-handling strategy.

Following this screening, CBF-ANFIS was retrained using the finalized experimental configuration, with a batch size of 256, a learning rate of 0.001, weight decay of 0.0001, a maximum of 100 epochs, and early-stopping patience of 12. Although both stages used the same data partitions, IndoBERT Mean representation, 12-rule ANFIS architecture, and experimental seeds, their training configurations differed. Consequently, the final CBF-ANFIS results (validation Macro-F1: 0.7048 ± 0.0040; test Macro-F1: 0.6897 ± 0.0054) should not be directly interpreted as evidence of a loss-function effect relative to the preliminary results. Table 5 therefore reports the preliminary configuration-selection results, while the controlled comparison of loss functions under the finalized configuration is presented in Section 4.8.

Table 5. Preliminary Comparison of Class-Imbalance Handling Strategies

Method	Validation Macro-F1	Test Macro-F1	ΔF1 vs Plain	Test Balanced Acc.	Test MCC
Plain CE	0.7081 ± 0.0058	0.6942 ± 0.0011	Reference	0.6901 ± 0.0007	0.5855 ± 0.0016
Class-balanced focal	0.7108 ± 0.0056	0.6954 ± 0.0045	+0.0013	0.6928 ± 0.0046	0.5867 ± 0.0060
Label smoothing	0.7101 ± 0.0076	0.6958 ± 0.0058	+0.0016	—	—
Logit adjustment	0.7093 ± 0.0069	0.6924 ± 0.0061	−0.0018	—	—

Note: Values are mean ± standard deviation across three seeds (42, 52, and 62). ΔF1 denotes the difference in test Macro-F1 relative to Plain ANFIS. CE, cross-entropy.

4.4. Effect of Metaheuristic Optimization

The final evaluation examined whether WOA and CSA could improve CBF-ANFIS performance by optimizing its antecedent parameters while keeping the consequent parameters fixed. Both algorithms were independently applied to the corresponding final CBF-ANFIS checkpoints using the same search configuration and validation Macro-F1 fitness function.

As shown in Table 6, the original CBF-ANFIS checkpoints achieved a mean validation Macro-F1 of 0.7048 ± 0.0040. WOA increased this value to 0.7111 ± 0.0017 (+0.0063), while CSA achieved 0.7074 ± 0.0039 (+0.0026). These results indicate that both optimizers improved the validation objective relative to their source checkpoints. However, these improvements did not translate into better held-out test performance. CBF-ANFIS achieved a test Macro-F1 of 0.6897 ± 0.0054, compared with 0.6890 ± 0.0089 for WOA-CBF-ANFIS and 0.6859 ± 0.0058 for CSA-CBF-ANFIS. The corresponding differences were −0.0007 and −0.0038, respectively. Furthermore, the validation-to-test gap increased from −0.0151 for CBF-ANFIS to −0.0221 for WOA-CBF-ANFIS and −0.0215 for CSA-CBF-ANFIS. Overall, although both optimizers improved validation Macro-F1, neither demonstrated a corresponding gain in predictive generalization.

Table 6. CBF-ANFIS Performance Before and After Metaheuristic Optimization

Model	Validation Macro-F1	Δ Val F1 vs CBF	Test Macro-F1	Δ Test F1 vs CBF	Val→Test Gap
CBF-ANFIS	0.7048 ± 0.0040	Reference	0.6897 ± 0.0054	Reference	−0.0151
WOA-CBF-ANFIS	0.7111 ± 0.0017	+0.0063	0.6890 ± 0.0089	−0.0007	−0.0221
CSA-CBF-ANFIS	0.7074 ± 0.0039	+0.0026	0.6859 ± 0.0058	−0.0038	−0.0215

Note: Values are mean ± standard deviation across three seeds (42, 52, and 62). Val–test gap = test Macro-F1 − validation Macro-F1.

4.5. Class-Wise Performance Analysis

Table 7 presents the class-wise performance of the final CBF-ANFIS rerun on the held-out test set. Crisis achieved the highest F1-score of 0.7721 ± 0.0011, with precision of 0.7745 ± 0.0073 and recall of 0.7699 ± 0.0092. Distress achieved an F1-score of 0.6887 ± 0.0040, followed by Situational at 0.6646 ± 0.0077.

Protective achieved the lowest F1-score of 0.6335 ± 0.0102, with precision of 0.6728 ± 0.0135 and recall of 0.5988 ± 0.0182. The lower recall indicates that Protective instances were more frequently misclassified than instances from the other categories. This pattern is examined further in Section 4.7.

Table 7. Class-Wise Performance of CBF-ANFIS

Class	Precision	Recall	F1-score
Distress	0.6591 ± 0.0089	0.7213 ± 0.0118	0.6887 ± 0.0040
Crisis	0.7745 ± 0.0073	0.7699 ± 0.0092	0.7721 ± 0.0011
Protective	0.6728 ± 0.0135	0.5988 ± 0.0182	0.6335 ± 0.0102
Situational	0.6696 ± 0.0102	0.6597 ± 0.0061	0.6646 ± 0.0077

4.6. Statistical Comparison

Paired bootstrap analysis with 1,000 resamples was performed on the final locked test predictions. As shown in Table 8, WOA−CBF produced ΔMacro-F1 values of −0.008888, +0.007296, and −0.000609 for seeds 42, 52, and 62, respectively. The 95% confidence interval excluded zero only for seed 52, indicating a seed-specific positive difference rather than a

consistent improvement across seeds. For CSA−CBF, the ΔMacro-F1 values were −0.007323, −0.000543, and −0.003671 for seeds 42, 52, and 62, respectively. The confidence interval excluded zero only for seed 42, indicating a seed-specific negative difference. Overall, the paired bootstrap analysis did not support a consistent improvement in held-out test performance from WOA or CSA optimization.

Table 8. Paired Bootstrap Comparison of ANFIS Models

Comparison	Seed	ΔMacro-F1	95% Bootstrap CI
WOA − CBF	42	−0.008888	[−0.017969, +0.000221]
WOA − CBF	52	+0.007296	[+0.001942, +0.013090]
WOA − CBF	62	−0.000609	[−0.008802, +0.007633]
CSA − CBF	42	−0.007323	[−0.013268, −0.001641]
CSA − CBF	52	−0.000543	[−0.005371, +0.003826]
CSA − CBF	62	−0.003671	[−0.008985, +0.001889]

4.7. Error Analysis

The confusion-matrix analysis revealed similar error patterns across CBF-ANFIS, WOA-CBF-ANFIS, and CSA-CBF-ANFIS. The most frequent misclassifications occurred between Situational and Distress, with additional errors involving Situational being predicted as Protective, Protective as Distress, and Crisis as Situational. These patterns were observed across the three random seeds, although their relative frequencies varied between models. Protective also exhibited the lowest recall among the four classes, indicating greater difficulty in correctly identifying this category.

Inter-model prediction agreement between the original CBF-ANFIS model and its metaheuristic-optimized variants was assessed using the same test instances for each corresponding seed. For WOA-CBF-ANFIS, the agreement rates were 91.66%, 96.88%, and 93.14% for seeds 42, 52, and 62, respectively. The corresponding rates for CSA-CBF-ANFIS were 96.47%, 97.94%, and 97.39%. These results indicate that antecedent optimization altered the predicted labels for a relatively small proportion of test instances while preserving substantial agreement with the original CBF-ANFIS predictions.

The distinction between Distress and Situational also presented challenges during annotation, with 1,240 disagreements involving these two categories. This observation is consistent with the frequent misclassifications between the categories, potentially reflecting the difficulty of distinguishing situational pressures from expressions of psychological distress. However, annotation disagreements were not directly matched with classification errors at the individual-post level; therefore, a causal relationship cannot be established. The error patterns and test performance indicate that antecedent optimization did not consistently improve generalization or resolve the classification difficulties observed across related categories. The contribution of class-balanced focal loss, however, cannot be isolated through this analysis. Its relative effect was therefore examined separately through the controlled loss-function ablation in Section 4.8.

4.8. Controlled Loss-Function Ablation

To assess the contribution of class-balanced focal loss (CBF) to the ANFIS classifier, a controlled ablation study compared CBF with three alternative objectives: standard cross-entropy (CE), label smoothing (LS), and logit adjustment (LA). All four loss functions were evaluated using the finalized experimental configuration, with the IndoBERT Mean 64D representation, 12-rule ANFIS architecture, fixed data partitions, core training hyperparameters, and three experimental seeds (42, 52, and 62) held constant. Only the loss function was varied to isolate its effect on classification performance. The original CBF-ANFIS checkpoints were also verified using SHA-256 integrity checks, and their validation and test performance was reproduced before conducting the comparisons.

As shown in Table 9, LS achieved the highest mean test Macro-F1 (0.6919), followed by LA (0.6902), CBF (0.6897), and CE (0.6892). The differences were small, with the largest reported improvement over CBF being approximately 0.0021. Although the alternative objectives achieved slightly higher mean validation Macro-F1 scores, these gains did not establish a consistent improvement in held-out test performance.

Table 9. Controlled loss-function ablation results across three random seeds

Loss function	Validation Macro-F1	Test Macro-F1	Δ Test Macro-F1 vs. CBF
CBF	0.7048 ± 0.0040	0.6897 ± 0.0054	—
CE	0.7081 ± 0.0024	0.6892 ± 0.0100	−0.0005
LS	0.7093 ± 0.0031	0.6919 ± 0.0069	+0.0021
LA	0.7092 ± 0.0006	0.6902 ± 0.0042	+0.0005

Note: Values are mean ± standard deviation across three seeds. Δ denotes the difference in test Macro-F1 relative to CBF.

The class-wise results further revealed differences in the relative performance of the loss functions. LS achieved the highest F1-scores for Crisis (0.7818) and Situational (0.6675), whereas LA performed best for Distress (0.6976) and Protective (0.6413). Protective remained the most challenging category across all loss functions, consistent with the findings in Section 4.5. These results suggest that incorporating class-balanced and focal weighting did not consistently improve the classification of individual categories relative to the alternative objectives.

Paired instance-level bootstrap analysis was conducted using 1,000 resamples for each seed, with comparisons based on predictions for the same test instances. All nine 95% percentile confidence intervals for the differences between CBF and the alternative objectives included zero. Thus, none of the seed-specific comparisons provided statistically distinguishable evidence of a performance difference. This finding should not, however, be interpreted as evidence of statistical equivalence between the methods. The controlled ablation shows that CBF achieved competitive performance but did not demonstrate consistent superiority over CE, LS, or LA. Its role is therefore better characterized as an imbalance-aware training strategy within the evaluated ANFIS framework rather than an empirically superior loss function.

4.9. Performance Differences

The magnitude of performance differences was examined using ΔMacro-F1, calculated from the mean held-out test Macro-F1 scores. As shown in Table 10, Direct IndoBERT achieved substantially higher Macro-F1 than TF-IDF + LR (+0.104009) and the ANFIS-based models (+0.074320 to +0.082626). These comparisons highlight the predictive performance gap between end-to-end contextual fine-tuning and the evaluated fixed-representation ANFIS approaches. However, differences involving preliminary Plain ANFIS are descriptive because it was trained under a different configuration from the final-stage models.

Within the ANFIS family, the final CBF-ANFIS model differed from preliminary Plain ANFIS by −0.004461. Relative to the final CBF-ANFIS baseline, WOA-ANFIS and CSA-ANFIS showed smaller decreases of −0.000734 and −0.003846, respectively. These differences indicate that antecedent optimization did not yield a measurable improvement in mean test Macro-F1. The controlled effects of loss-function selection and metaheuristic optimization are examined separately in Sections 4.8 and 4.4, respectively.

Table 10. Macro-F1 Differences between Models

Contrast	ΔMacro-F1
Direct IndoBERT − TF-IDF + LR	+0.104009
Direct IndoBERT − Plain ANFIS	+0.074320
Direct IndoBERT − CBF-ANFIS	+0.078781
Direct IndoBERT − WOA-ANFIS	+0.079515
Direct IndoBERT − CSA-ANFIS	+0.082626
CBF-ANFIS − Plain ANFIS	−0.004461
WOA-ANFIS − CBF-ANFIS	−0.000734
CSA-ANFIS − CBF-ANFIS	−0.003846

Note: ΔMacro-F1 represents the difference between the mean test Macro-F1 values of the two models in each contrast. Positive values indicate a higher Macro-F1 for the first model.

4.10. Rule-Level Interpretability Analysis

The rule-level analysis examined the final CBF-ANFIS checkpoints from three random seeds (42, 52, and 62) using 2,915 test samples. Each model comprised 12 fuzzy rules operating on 64-dimensional IndoBERT Mean representations. Prediction reproduction was verified for all three checkpoints, with identical test predictions obtained after reloading the saved models.

Each fuzzy rule defines a region in the latent feature space through its antecedent membership functions. Its normalized firing strength indicates the degree to which an input representation activates that rule, providing a traceable view of the model's inference process. Unlike a conventional classification output, this representation allows the contribution of individual rules to be examined before the final class decision is produced. However, these activation patterns describe relationships in the latent feature space rather than the semantic meaning of the original text.

Table 11. Rule Activation Analysis of CBF-ANFIS

Seed	Mean active rules	Mean maximum rule activation	Mean rule entropy	Prediction reproduction (%)
42	1.4724	0.8569	0.3930	100
52	1.4837	0.8446	0.4300	100
62	1.5276	0.8405	0.4229	100

As summarized in Table 11, the mean maximum rule activation ranged from 0.8405 to 0.8569 across the three seeds, while mean rule entropy ranged from 0.3930 to 0.4300. The mean number of active rules was relatively stable, ranging from 1.4724 to 1.5276 out of 12 available rules. Taken together, these values suggest that the inference process was generally concentrated on a limited subset of rules rather than being distributed uniformly across the rule base. The relatively high maximum activation and low number of active rules indicate a selective activation pattern, although they do not establish that individual rules represent distinct or semantically coherent categories.

Beyond activation strength, the contribution of each rule to the classification output was examined using its normalized firing strength and class-specific consequent output. This distinction is important because a strongly activated rule does not necessarily contribute positively or dominate the prediction for a particular class. Its actual contribution also depends on the magnitude and direction of its consequent parameters. Accordingly, examining activation and consequent outputs together provides a more informative account of rule-level inference than activation strength alone.

Despite this structural transparency, the 64 input features encode contextual information in a latent space rather than directly interpretable linguistic attributes. Individual fuzzy rules therefore cannot be directly associated with specific expressions of Distress, Crisis, Protective, or Situational conditions. Furthermore, rule indices are specific to each trained model and should not be assumed to represent equivalent patterns across different random seeds. Consequently, the analysis supports tracing how the model processes latent representations through its fuzzy rules, but does not establish direct semantic explanations of the original social-media posts.

4.11. Main Findings

The results highlight a distinction between predictive performance and the structural interpretability offered by the evaluated modeling approaches. Direct IndoBERT achieved the strongest predictive performance, outperforming the fixed-representation ANFIS models. This suggests that end-to-end contextual fine-tuning was more effective for capturing task-relevant information than using fixed text representations as inputs to the ANFIS architecture.

Within the ANFIS framework, the choice of text representation influenced classification performance, with IndoBERT Mean providing the most effective input representation among those evaluated. However, the improvements associated with class-imbalance handling were relatively modest, and the controlled comparisons did not establish a clear predictive advantage for a particular loss function. Similarly, although WOA and CSA improved validation

Macro-F1, these gains did not translate into better test performance. The paired bootstrap analysis further indicated that the observed differences were not statistically distinguishable, suggesting that antecedent optimization did not provide a demonstrable generalization benefit under the evaluated settings.

Taken together, these findings suggest that the main value of the ANFIS-based approach lies not in achieving higher predictive performance than direct IndoBERT, but in providing a structured framework for examining fuzzy-rule activation and parameter optimization alongside different text representations and training objectives. Thus, predictive effectiveness and rule-based model analysis emerge as distinct considerations in evaluating the proposed framework.

5. Comparison with Previous Studies

Table 12 positions the present study within previous research on mental-health text classification, highlighting differences in datasets, text representations, classification methods, and interpretability. Earlier studies have explored a range of approaches, from handcrafted linguistic and statistical features to word embeddings and contextual Transformer representations. However, their reported results were obtained under different datasets, languages, class definitions, sampling procedures, and evaluation protocols. Consequently, the performance values provide methodological context rather than a basis for direct numerical comparison.

Table 12. Comparison with Previous Mental-Health NLP Studies

Study	Dataset	Representation	Method	Classes	Accuracy / F1	Interpretability
Tadesse et al. [7]	Reddit	LIWC + LDA + bigram	MLP	2	Acc. 0.91; F1 0.93	Feature-level analysis; no intrinsic rules
Figuerêdo et al. [8]	eRisk 2017 (Reddit)	Word embeddings + emotion-related features	CNN with early/late fusion	2	F1 0.71	Emotion-feature analysis; no intrinsic rules
Kerasiotis et al. [9]	Social media depression dataset	DistilBERT + metadata + linguistic features	Transformer + MLP	4	Weighted F1 0.8415	No intrinsic rule-based interpretation
Present study	Indonesian social media	IndoBERT Mean	WOA–ANFIS	4	Acc. 0.6883 ± 0.0088; Macro-F1 0.6890 ± 0.0089	Latent-space fuzzy-rule activation

The studies illustrate different strategies for extracting mental-health-related information from social-media text. Tadesse et al. [7] combined linguistic and statistical features with an MLP classifier for binary depression detection, while Figuerêdo et al. [8] incorporated emotion-related information into an embedding-based CNN framework for early depression detection. Kerasiotis et al. [9] extended this direction by combining DistilBERT representations with metadata and linguistic features for four-class depression-severity classification. These approaches reflect a progression from manually designed features toward contextual representations and multimodal or auxiliary feature integration. Nevertheless, predictive performance alone does not indicate whether the resulting models provide an explicit representation of their inference processes.

The present study investigates a different combination of components by integrating contextual text representation with neuro-fuzzy classification, class-imbalance handling, and metaheuristic antecedent optimization. Under the same fixed data partitions, direct IndoBERT achieved the highest Macro-F1 of 0.7685 ± 0.0046, whereas the ANFIS-based models achieved lower scores, ranging from 0.6859 to 0.6942. This result indicates that end-to-end contextual fine-tuning was more effective for predictive classification than the evaluated fixed-representation ANFIS configurations. The contribution of the ANFIS framework therefore lies not in outperforming direct Transformer fine-tuning, but in providing an explicit rule-based inference structure through which activation patterns and consequent contributions can be examined.

An important distinction concerns the level of interpretability provided by these approaches. The previous studies summarized in Table 12 primarily rely on feature-level analysis, auxiliary linguistic information, or conventional neural classification structures. In contrast, ANFIS explicitly represents its inference process through fuzzy-rule activation and

consequent parameters. However, because its inputs are 64-dimensional IndoBERT latent representations, these rules describe relationships in the reduced feature space rather than directly interpretable linguistic patterns. Thus, the framework provides structural transparency at the rule level, but does not establish direct semantic explanations of individual social-media posts.

Overall, the comparison positions the present study as an investigation of the relationship between contextual representation, neuro-fuzzy inference, and predictive performance in four-class Indonesian mental-health indicator classification. Its findings also highlight an important distinction: stronger predictive performance and more explicit rule-level inference are separate modeling objectives and should not be treated as interchangeable measures of model quality.

6. Conclusions

This study investigated the integration of contextual text representation, neuro-fuzzy classification, class-imbalance handling, and metaheuristic antecedent optimization for Indonesian social-media mental-health indicator classification. The experiments showed that the effectiveness of these components varied across the evaluated settings. Direct IndoBERT fine-tuning achieved the strongest predictive performance, while IndoBERT Mean provided the most effective fixed representation for ANFIS. However, neither class-balanced focal loss nor WOA and CSA optimization demonstrated consistent improvements in held-out test performance. These findings address the study's objective by showing that contextual representation was the strongest determinant of predictive performance, whereas the additional ANFIS and optimization components did not establish a comparable predictive advantage.

The contribution of the ANFIS-based approach lies in its explicit inference structure, which allows rule activation and consequent contributions to be examined within the reduced latent feature space. This provides a means of tracing model behavior beyond the final classification output. Nevertheless, such structural transparency should not be equated with direct semantic interpretability, since the input features do not correspond to directly identifiable words, phrases, or psychological attributes. The framework should therefore be understood as a structured classification approach for analyzing mental-health-related indicators, rather than a clinical assessment or diagnostic system.

Several limitations should be considered when interpreting these findings. First, the dataset was collected from X using predefined retrieval keywords, which may restrict the coverage of expressions and experiences represented in the resulting corpus. Second, although the annotation process included independent labeling and consensus adjudication, the annotators were not clinical mental-health professionals. The resulting labels therefore represent text-based category assignments rather than clinically validated assessments. Third, the use of 64-dimensional latent representations limits the extent to which fuzzy rules can be translated into meaningful linguistic or psychological explanations. In addition, the evaluation was conducted on a single Indonesian social-media dataset, and the observed performance has not yet been established across independent datasets or other language and domain settings.

Future work should investigate the framework using independently collected and externally validated datasets, broader linguistic and domain settings, and annotation protocols involving qualified mental-health professionals. Further research could also explore more semantically grounded representations, calibrated predictions, and complementary explanation methods that connect latent-space rule behavior with interpretable textual evidence. These directions are necessary to determine whether the framework can provide more reliable and generalizable support for mental-health-related text classification.

Author Contributions: Conceptualization: H.B and S.S.B; Methodology: H.B and S.S.B; Software: S.S.B; Validation: H.B, and A.R.H.; Formal analysis: H.B, S.S.B, and A.R.H; Resources: H.B; Data curation: F.Y.; Writing—original draft preparation: H.B, S.S.B, and A.R.H; Writing—review and editing: M.R, and F.Y.; Supervision: H.B; funding acquisition: H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Directorate of Research and Community Service (DPPM), Ministry of Higher Education, Science, and Technology of the Republic of

Indonesia, under Grant No. 0350.09/LL5-INT/AL.04.03/2026, 005/S.Kontrak-LIT./LPPM/V/2026.

Data Availability Statement: A GitHub repository associated with this study is provided at https://github.com/fayfourteen/MentalHealth-NeuroFuzzy-Methaheuristic. The repository is intended to support research transparency and reproducibility. Due to the sensitive nature of mental-health-related social-media content and applicable data-sharing restrictions, the availability of original posts and annotated records is subject to privacy and platform requirements. Additional research materials may be requested from the corresponding author, subject to applicable ethical and data-sharing restrictions.

Acknowledgments: The authors acknowledge the use of generative artificial intelligence (AI) tools to assist with literature discovery, initial drafting and language refinement of the manuscript, and troubleshooting programming-code errors encountered during the research process. The AI tools were used only as supporting tools and were not involved in the formulation of the research methodology, interpretation of the results, or scientific decision-making. All references suggested or identified with AI assistance were independently verified by the authors against the original sources. The authors reviewed, revised, and validated all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digit. Med.*, vol. 5, no. 1, p. 46, Apr. 2022, doi: 10.1038/s41746-022-00589-7.

[2] R. Salas-Zárate, G. Alor-Hernández, M. del P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, "Detecting Depression Signs on Social Media: A Systematic Literature Review," *Healthcare*, vol. 10, no. 2, p. 291, Feb. 2022, doi: 10.3390/healthcare10020291.

[3] A. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review," *Appl. Soft Comput.*, vol. 130, p. 109713, Nov. 2022, doi: 10.1016/j.asoc.2022.109713.

[4] K. Mao, Y. Wu, and J. Chen, "A systematic review on automated clinical depression diagnosis," *npj Ment. Heal. Res.*, vol. 2, no. 1, p. 20, Nov. 2023, doi: 10.1038/s44184-023-00040-z.

[5] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.

[6] G. Huang, Y. Li, S. Jameel, Y. Long, and G. Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?," *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 362–373, Dec. 2024, doi: 10.1016/j.csbj.2024.05.004.

[7] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[8] J. S. L. Figuerêdo, A. L. L. M. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," *Online Soc. Networks Media*, vol. 31, p. 100225, Sep. 2022, doi: 10.1016/j.osnem.2022.100225.

[9] M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 196, Sep. 2024, doi: 10.1007/s13278-024-01360-4.

[10] K. Yang, T. Zhang, and S. Ananiadou, "A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media," *Inf. Process. Manag.*, vol. 59, no. 4, p. 102961, Jul. 2022, doi: 10.1016/j.ipm.2022.102961.

[11] E. Bao, A. Pérez, and J. Parapar, "Explainable depression symptom detection in social media," *Heal. Inf. Sci. Syst.*, vol. 12, no. 1, p. 47, Sep. 2024, doi: 10.1007/s13755-024-00303-9.

[12] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Min. Knowl. Discov.*, vol. 11, no. 5, p. e1424, Sep. 2021, doi: 10.1002/widm.1424.

[13] X. Gu, J. Han, Q. Shen, and P. P. Angelov, "Autonomous learning for fuzzy systems: a review," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 7549–7595, Aug. 2023, doi: 10.1007/s10462-022-10355-6.

[14] L. Pickering, K. Cohen, and B. De Baets, "A Narrative Review on the Interpretability of Fuzzy Rule-Based Models from a Modern Interpretable Machine Learning Perspective," *Int. J. Fuzzy Syst.*, vol. 28, no. 4, pp. 1313–1332, Jun. 2026, doi: 10.1007/s40815-025-02022-z.

[15] M. Mehrabi, B. Pradhan, H. Moayedi, and A. Alamri, "Optimizing an Adaptive Neuro-Fuzzy Inference System for Spatial Prediction of Landslide Susceptibility Using Four State-of-the-art Metaheuristic Techniques," *Sensors*, vol. 20, no. 6, p. 1723, Mar. 2020, doi: 10.3390/s20061723.

[16] I. Ahmadianfar, S. Shirvani-Hosseini, J. He, A. Samadi-Koucheksaraee, and Z. M. Yaseen, "An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction," *Sci. Rep.*, vol. 12, no. 1, p. 4934, Mar. 2022, doi: 10.1038/s41598-022-08875-w.

[17] H. Wang, B. Chen, H. Sun, A. Li, and C. Zhou, "AnFiS-MoH: Systematic exploration of hybrid ANFIS frameworks via metaheuristic optimization hybridization with evolutionary and swarm-based algorithms," *Appl. Soft Comput.*, vol. 167, p. 112334, Dec. 2024, doi: 10.1016/j.asoc.2024.112334.

[18] L. Deng and S. Liu, "Deficiencies of the whale optimization algorithm and its validation method," *Expert Syst. Appl.*, vol. 237, p. 121544, Mar. 2024, doi: 10.1016/j.eswa.2023.121544.

[19] A. G. Hussien *et al.*, "Crow Search Algorithm: Theory, Recent Advances, and Applications," *IEEE Access*, vol. 8, pp. 173548–173565, 2020, doi: 10.1109/ACCESS.2020.3024108.

[20] A. Sheta, M. Braik, H. Al-Hiary, and S. Mirjalili, "Improved versions of crow search algorithm for solving global numerical optimization problems," *Appl. Intell.*, vol. 53, no. 22, pp. 26840–26884, Nov. 2023, doi: 10.1007/s10489-023-04732-z.

[21] R. Budianoor, S. W. Saputro, F. Abadi, R. A. Nugroho, and A. Farmadi, "Quantifying the Impact of Text Preprocessing on IndoBERT Fine-Tuning for Indonesian Informal Culinary Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 564–581, May 2026, doi: 10.62411/jcta.15980.

[22] I. N. Aqiilah, R. Saptono, and A. Syaifuddin, "Sentence-Level Sentiment Analysis of Indonesian App Reviews Using IndoBERTweet," *J. Comput. Theor. Appl.*, vol. 4, no. 1, pp. 74–91, Jun. 2026, doi: 10.62411/jcta.16240.

[23] B. R. P. Darnoto and D. B. Firmawan, "Language-Similarity-Guided Transfer Fine-Tuning of Pre-trained Transformer Models for Sentiment Analysis Across 12 Indonesian Regional Languages," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 547–563, May 2026, doi: 10.62411/jcta.15975.

[24] S. Mahmood, N. Z. Bawany, and M. R. Tanweer, "A comprehensive survey of whale optimization algorithm: modifications and classification," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 29, no. 2, p. 899, Feb. 2023, doi: 10.11591/ijeecs.v29.i2.pp899-910.

[25] T. Thaher, A. Sheta, M. Awad, and M. Aldasht, "Enhanced variants of crow search algorithm boosted with cooperative based island model for global optimization," *Expert Syst. Appl.*, vol. 238, p. 121712, Mar. 2024, doi: 10.1016/j.eswa.2023.121712.

[26] M. Braik, H. Al-Zoubi, M. Ryalat, A. Sheta, and O. Alzubi, "Memory based hybrid crow search algorithm for solving numerical and constrained global optimization problems," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 27–99, Jan. 2023, doi: 10.1007/s10462-022-10164-x.

[27] A. D. Latief, A. Jarin, M. T. Uliniansyah, E. Nurfadhilah, and D. I. N. Afra, "A Proven Sentiment Annotation Guideline for Indonesian Twitter Data," in *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Oct. 2023, pp. 31–36. doi: 10.1109/IC3INA60834.2023.10285807.

[28] B. M. Rashed and N. Popescu, "Medical Image-Based Diagnosis Using a Hybrid Adaptive Neuro-Fuzzy Inferences System (ANFIS) Optimized by GA with a Deep Network Model for Features Extraction," *Mathematics*, vol. 12, no. 5, p. 633, Feb. 2024, doi: 10.3390/math12050633.

[29] N. Das and S. A. Begum, "An Empirical Study of Loss Functions for Aspect Category Detection in Imbalanced Data Scenario," in *2025 10th International Conference on Signal Processing and Communication (ICSC)*, Feb. 2025, pp. 247–252. doi: 10.1109/ICSC64553.2025.10968230.

[30] H. Ren, Y. Zhao, Y. Zhang, and W. Sun, "Learning label smoothing for text classification," *PeerJ Comput. Sci.*, vol. 10, p. e2005, Apr. 2024, doi: 10.7717/peerj-cs.2005.

[31] H. Lee, T. Park, and H. Kim, "Learnable Logit Adjustment for Imbalanced Semi-Supervised Learning Under Class Distribution Mismatch," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2025, pp. 2664–2674. doi: 10.1109/ICCV51701.2025.00256.

[32] M. H. Nadimi-Shahraki, H. Zamani, Z. Asghari Varzaneh, and S. Mirjalili, "A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations," *Arch. Comput. Methods Eng.*, vol. 30, no. 7, pp. 4113–4159, Sep. 2023, doi: 10.1007/s11831-023-09928-7.

[33] Y. Meraihi, A. B. Gabis, A. Ramdane-Cherif, and D. Acheli, "A comprehensive survey of Crow Search Algorithm and its applications," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 2669–2716, Apr. 2021, doi: 10.1007/s10462-020-09911-9.

[34] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, Jun. 2024, doi: 10.1162/tacl_a_00675.

[35] M. Carpita and S. Golia, "Categorical classifiers in multiclass classification with imbalanced datasets," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 16, no. 4, pp. 391–405, Aug. 2023, doi: 10.1002/sam.11624.

[36] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.