SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I5P122

Volume 13 Issue 5, 272-285, May 2026

Original Article

Advancing Fake News Detection: Multimodal LSTM Optimized by Grey Wolf Algorithm with Explainable AI

Raed Abdul Karim Al-Jabri[1], Mohsen Rezvani[2]

[1,2]*Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran*.

[2]*Corresponding Author : mrezvani@shahroodut.ac.ir*

Received: 18 February 2026 Revised: 18 March 2026 Accepted: 21 April 2026 Published: 27 May 2026

Abstract *- Due to the social media explosion, the spread of misinformation is increasing, and it presents a severe social and political threat. These applications are typical of many fake news detection systems that rely solely on textual data and employ black-box models, which limit robustness and interpretability. In this paper, we introduce a new multimodal model that combines textual semantics, user behavior, and Propagation properties together in a unified framework. The semantic representation is learned using FastText embeddings, and the patterns of diffusion are represented by propagation-based features. An LSTM network based on the GWO algorithm achieves good classification for stable data. Shapley Additive Explanations (SHAP) is used to explain the model predictions and improve interpretability. We conduct empirical evaluation on Multi-Fake-DetectiVE 2023 and TAGFN, where our model matches the state-of-the-art performance with up to 99.32% accuracy in non-adversarial settings and outperforms other techniques by several orders of magnitude under adversarial setups. Concluding Remarks: We have demonstrated that the combination of multimodal learning, evolutionary optimization, and explainable AI guarantees strong amounts of robustness in detecting fake news.*

Keywords *- Fake News Detection, Multimodal Learning, Fasttext Embeddings, Long Short-Term Memory (Lstm), Grey Wolf Optimizer (GWO), SHAP.*

1. Introduction

The rapid expansion of digital communication technologies has significantly transformed the production and consumption of information. Social media platforms have become the primary source of news for millions of users worldwide, often replacing traditional news outlets in many contexts [1, 2]. While this transformation has enhanced accessibility and speed of information dissemination, it has also created an environment where misinformation and disinformation can spread rapidly with minimal verification.

Fake news, defined as intentionally misleading or fabricated information presented as legitimate news, has emerged as a serious societal challenge [3, 4]. Its impact extends beyond individual misinformation to influence political processes, economic decisions, and public trust in institutions [5]. Large-scale misinformation campaigns, particularly during critical political events, demonstrate the ability of manipulated content to shape public opinion and intensify polarization [6]. Furthermore, empirical studies reveal that false information tends to spread faster and attract more engagement than factual content due to emotional appeal and algorithmic amplification mechanisms [7–9]. Given the massive volume of online data generated daily, manual verification is infeasible, making automated fake news detection systems an essential research focus [10–12]. Early approaches relied on traditional machine learning techniques using handcrafted textual features and statistical models [13]. However, these methods are limited in capturing deep semantic relationships and contextual dependencies in complex information.

More recently, deep learning and transformer-based models have achieved significant improvements in fake news detection by learning rich representations directly from raw textual data [14]. Despite these advancements, existing approaches still suffer from several important limitations. First, most models are predominantly text-centric and fail to incorporate behavioral and propagation-based features, such as user interactions and information diffusion patterns, which play a critical role in the spread of misinformation [15].

Secondly, complex Deep Learning models work as black-box systems, which have limited interpretability, and thus diminish the transparency and trust of their predictions in practical applications throughout the world [16]. Such concessions also reveal a grim research gap, namely, the lack of an integrated framework that integrates semantic content, user behavior, and propagation dynamics jointly while being robust as well as interpretable. Addressing this gap is

necessary to establish and design efficient and usable fields of fake news detection systems.

But in practice, the diffusion of fake news is not grounded simply in text and conditional on structural and behavioral variables. If we closely observe those complementary patterns across user interaction, temporal diffusion structures, and engagement dynamics, we can significantly improve the performance of detection. Therefore, there is a high need for detection systems that integrate semantic, behavior, and propagation-based features to create a single learning framework.

Motivated by these challenges, this work proposes a fraud news detection framework based on multimedia textual indexes, user-behavior characteristics, and propagation models. Temporal dependencies are extracted by utilizing a Long-Short Term Memory (LSTM) network, while to automatically tune hyperparameters, Grey Wolf Optimizer (GWO) is implemented, which increases convergence stability and generalization performance.

The method also includes Shapley Additive Explanations (SHAP), further interpreting results, making the model predictions more transparent by explaining individual features' contributions, leading towards clearer conclusions.

The primary contributions of this research are summarized as follows:

- Development of a unified multimodal detection architecture integrating semantic, behavioral, and diffusion-aware features.
- Application of evolutionary optimization using GWO to enhance model robustness and convergence efficiency.
- Evaluation of model performance under adversarial manipulation scenarios to assess robustness.
- Integration of explainable AI techniques using SHAP to improve transparency and interpretability of model decisions.

The rest of this paper is organized as follows. Section 2 summarizes related work. Methodology: The methodology is further detailed in section three. Section 4 presents the experimental results and analysis. Section 5 concludes the study and outlines areas for future research.

2. Related Work

In recent years, fake news detection has witnessed an evolution from pipeline-based traditional machine learning to context-sensitive neural models and evidence cognizant graph frameworks. Most of the existing research can be classified into four large directions: distributed and scalable machine learning, evolutionary optimization of classifiers, representation-centric deep learning models, and evidence modeling based on graphs.

2.1. Distributed and Scalable Machine Learning

Fake news detection is also a scalability issue due to the huge scale and speed at which data is identified on social media. Mmm, e.g., big data = distributed ML frameworks? Methods: A taxonomy-based approach employs large-scale ensemble classifiers populated with word n-grams and TF-IDF-based bag-of-words representations trained using big-data infrastructures [17]. Computational efficiency and better performance in terms of F1-score, these methods are implemented on distributed computing platforms (eg, Apache Spark).

That said, many of these approaches are scalable but also depend on an engineered set of features and a traditional classifier. This limits their ability to generalize to novel linguistic trends or adversarial instances.

2.2. Evolutionary Optimization of Classical Models

Fake news detection pipelines have incorporated evolutionary algorithms to improve classification performance while maintaining complex-intensity models. Several dataset-specific techniques shall find their best configuration of classifiers using Genetic Algorithm (GA) based methods, such as Support Vector Machines, Naïve Bayes, Random Forest, and Logistic Regression [18]. These methods offer improved accuracy with quicker convergence to the solution.

Still, those approaches are text-heavy and dependent on fairly surface-level lexical features. Specifically, they are not semantically context-specific and temporal dynamical and Propagation behavioral that is essential for the study of misinformation in real-world care.

2.3. Application-Specific and Behavioral Modeling

Work has also been done with respect to fake news detection for specific application domains aside from general news classification. For instance, AI-based platforms have been suggested to neutralize misinformation in supply chain systems by integrating multi-source data signals against production and consumption noise affecting productive decision-making [19].

These ones, and other similar techniques, are good for use in domain-specific basic settings, but tend not to be generalizable to broader social media. It also includes the research stream of swimming through user behavior modeling and interaction patterns.

Fake or bot accounts have been identified using follower counts, activity patterns, and engagement metrics-based features through a Neural Network [20]. While these methods improve the systemic understanding of how information spreads and the impact of users, they mainly target classification at the user level but leave out knowledge about the news content itself with regard to its credibility.

2.4. Representation-Centered Deep Learning Models

However, Deep learning has provided a quantum leap in fake news detection since it allows us to learn the semantic representation of our text. It has been shown that more complex classifiers may or may not improve classification performance when compared to simpler models, especially when used in conjunction with document embedding techniques, which can provide much more contextual encoding [21].

Finally, the transformer-based models, inspired by BERT [9], improved parameterization capabilities for contextual understanding in localized environments like social media [22].

It is also proposed that hybrid architectures, where global contextual representations can be fused with local feature extraction, can be used. Previous works (such as BERT–CNN models) combine contextual relations and n-gram-level patterns to boost performance [23]. Various multiscale transformers and hierarchical architectures have been explored for a multilingual and multi-contextual approach [24, 25] using code-mixed and cross-lingual data.

However, rule-based methods like Google Factbook or machine learning approaches with a single contextual encoder perform poorly on domain-specific cases (e.g., COVID-19 misinformation detection [26]) compared to fusion-based models that utilize multiple contextual encoders.

These are recent advancements of these methods and have been taken a step further by fusing multi-modal features embedding and large language models. We introduce a knowledge-enriched benchmark, FineFake, containing multi-domain and fine-grained annotations plus multimodal content, and show that existing models see huge performance drops in cross-domain scenarios [17].

In addition, multimodal fusion frameworks such as hierarchical progressive transformers blend text, vision, and contextual information with knowledge derived from LLMs to increase performance and Explainability [18].

Furthermore, recent work has noted robustness challenges. State-of-the-art detection methods can be attacked using very small perturbations in sentiment in a new adversarial framework [22]. Common frameworks like FACTGUARD also utilize large language models for the task of event-centric information extraction, while reducing the reliance on writing style features [25], making it more robust and reliable.

However, almost all deep learning-based methods are content-centric and poorly model propagation dynamics, robustness, and interpretability. Table 1 is a summary of the comparison of these methods.

2.5. Evidence-Aware Graph Neural Network Frameworks

Related works: Two major implementation changes in fake news detection occurred due to the introduction of Graph Neural Networks (GNNs) that model the relationships between claims and supporting evidence. These methods enhance semantic reasoning and information aggregation by modeling news data and the interactions between users and news content as graphs [27].

More recently, it has been combined with contrastive learning and adversarial training methods to improve the robustness of graph representations [28]. More sophisticated methods have additionally added multimodal data and modes of interaction into a variety of graph-based models.

Local Dynamic Propagation Graph (LDPG): LDPG [28] is a relatively new design, which aims to understand the temporal dynamics of user-user information diffusion. This strategy improves recognition robustness capabilities and initial prediction performance through jointly exploiting multimodal attributes of the content along with time-sensitive graph representations. We contrast these advancements with established methods (Table 1).

While they generally perform well, graph-based approaches add another layer of computation and also need structured propagation data to train on—a representation that may not always be available in practice.

While attempting to combine Explainability and multimodal fusion into a single architecture has great potential, the approach is also ambitious with untapped pitfalls.

Despite significant improvements in fake news detection regarding Critical Thinking-based FaKP, some limitations remain. Current methods are mainly content-centric and do not fully exploit semantic, behavioral, or propagation-based information.

Moreover, many deep learning approaches are black boxes by nature, with low interpretability and trustworthiness of predictions. In addition, adversarial robustness and cross-domain generalization remain under-addressed.

Additionally, most current models perform worse than one would expect in the real world since they cannot generalize to changing or domain-specific misinformation patterns.

So, there is an urgent challenge to design a unified architecture where multimodal features, including textual information, behavioral information, and propagation info, are integrated with robust and interpretable ability. This gap enforces the motivation behind the proposed approach in this study.

Table 1. Comparative Summary of Representative Fake News Detection Studies

Ref.	Dataset / Setting	Feature Set	Method	Reported Results
[17]	FineFake (multi-domain, multi-platform)	Multimodal content + social context + external knowledge + fine-grained annotations	Knowledge-enhanced benchmark for cross-domain detection	F1 drops 10–30% (increased task difficulty)
[18]	WeiBo21, FineFake	Text + image + LLM knowledge + user comments	LLM-MFEFND (Hierarchical Progressive Transformer + multimodal fusion + explainability)	Acc: 94.5%, 81.1%; +1.7–2.2% improvement
[19]	Multi-source supply chain cases	Multi-source AI/ML signals	AI/ML-driven FNaD detection	Effective domain-specific performance
[20]	Twitter + synthetic/real datasets	Behavioral/user attributes	RNN-based account classification	Acc: 96% (real), 98% (synthetic)
[21]	Five news corpora	Document embeddings	Embedding-based classifiers	Competitive accuracy
[22]	Multiple benchmark datasets	Contextual embeddings + sentiment-aware features	AdSent (adversarial sentiment-robust framework)	Improved robustness & generalization
[23]	Kaggle dataset	Global + local text semantics	BERT–CNN hybrid	~1.10% improvement over baseline
[24]	Mixed-language dataset	Multiscale contextual features	Multiscale Transformer	2–10% improvement
[25]	Benchmark datasets	Event-centric semantic + LLM reasoning	FACTGUARD (LLM-guided reasoning + knowledge distillation)	Higher robustness & accuracy
[26]	CONSTRAINT-2021 (COVID misinformation)	Contextual embedding fusion	Early fusion (BERT/XLNet/ELMo)	Acc: 97%
[27]	Evidence-aware datasets	Graph-structured semantic features	GETRAL (GNN + Contrastive + Adversarial learning)	Outperforms prior SOTA
[28]	PHEME, PolitiFact, GossipCop	Multimodal + temporal propagation graph features	LDPG (Dynamic GNN + attention + modality alignment)	Superior performance & early detection robustness
Proposed Model	Multi-Fake-DetectiVE 2023, TAGFN	Text + User + Propagation	GWO-Optimized LSTM + SHAP	Acc: 99.32%, RI: 0.94

3. Methodology

As illustrated in Figure 1, the pipeline consists of collecting multimodal fake news data from texts and the propagation graph of TAGFNs, pre-processing/feature extraction on those data, training a GWO-optimized LSTM model, and explaining it with SHAP.

This fully integrated end-to-end pipeline allows for modeling the semantic content, user behavior, and propagation dynamics coherently in a form that can be reliably detected. These components are discussed in the following subsections.

Furthermore, the integration of optimized deep learning with structured propagation features enables the model to capture both temporal dependencies and relational interactions effectively. The GWO-based optimization mechanism ensures adaptive tuning of model parameters, thereby improving convergence speed and generalization capability. In addition, the inclusion of SHAP-based Explainability provides transparent insights into feature contributions, enhancing model interpretability and trustworthiness. Overall, the proposed framework offers a robust and scalable solution for real-world fake news detection scenarios.

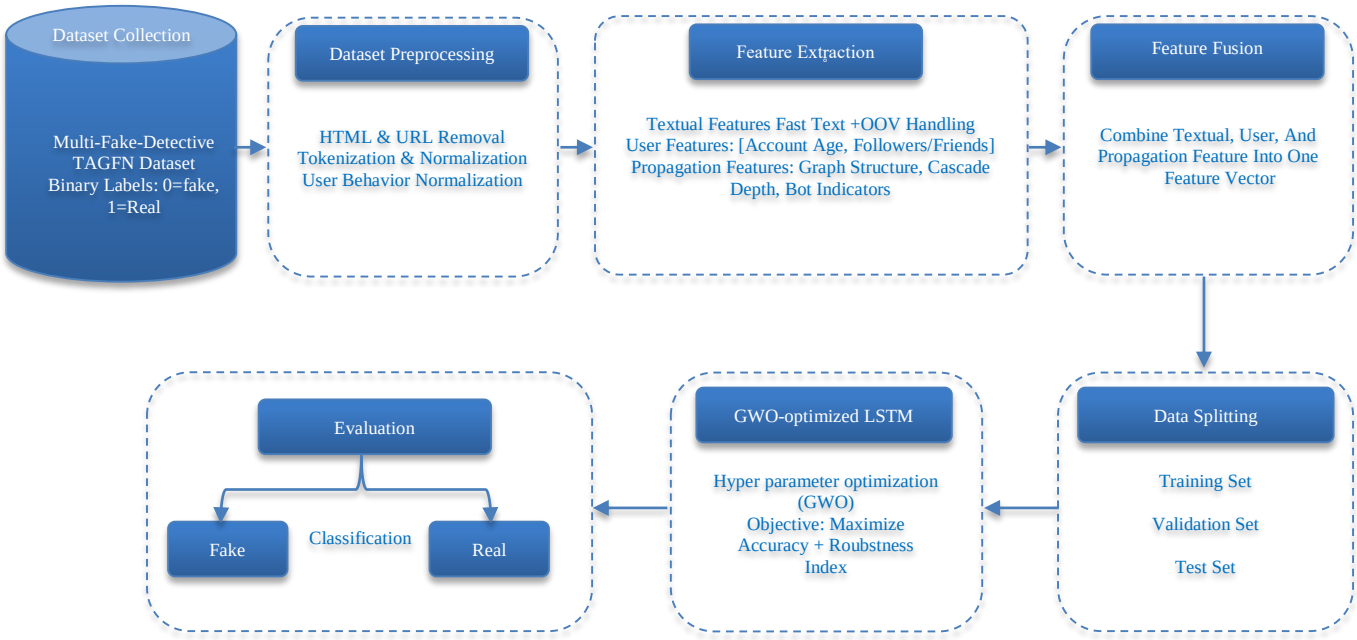


Fig. 1 Overview of the proposed multimodal fake news detection framework

3.1. Dataset

In this work, we cast the fake news detection problem into a binary classification where label 0 indicates that the news is fake and label 1 indicates that the news is real. The experiments are conducted on two public datasets, Multi-Fake-DetectiVE 2023 [29] and TAGFN – Text-Attributed Graph Dataset for Fake News Detection [30]. The statistical distribution of the two datasets is presented in Table 2.

To enable fair comparison, we divide both datasets into training, validation, and testing sets with a stratified sampling approach. This property ensures the integrity of class distribution and the minimization of sampling bias in various experimental runs.

Table 2. Statistical distribution of the employed datasets

Dataset	Fake	Real	Total
Multi-Fake-DetectiVE 2023 [29]	35,910	35,940	71,850
TAGFN Dataset [30]	14,170	14,130	28,300

3.2. Data Preprocessing

The role of the text preprocessing is to remove noise and determine a representation for the input. Text is preprocessed such that it's all lower-cased and non- ASCII characters are removed prior to tokenization. Stop-words are removed, and Python regular expressions (RegEx) are utilized to remove URLs, punctuation marks, numerical tokens, emails, and contact numbers. Duplicate data is also removed, as it may introduce bias into the training of models [31]. A possible list of pre-processing tasks is shown in Table 3. Furthermore, lemmatization is applied to reduce words to their base forms, thereby improving semantic consistency across the corpus.

Finally, the processed text is converted into numerical representations using tokenization and padding techniques to ensure uniform input length for the deep learning model.

Table 3. Examples of original and preprocessed news text

ID	Original Text	Cleaned Text
1	Breaking News!! COVID-19 cure found today!!!	Breaking news covid cure has been found today.
2	Donald Trump shared fake info via Twitter!!!	Donald Trump shares fake info on Twitter

3.3. Feature Extraction

More precisely, the pre-processed emerges as discriminative numerical vectors in a multimodal perspective that captures semantic and Propagation properties of fake news, also see Figure 2. Text features: The obtained text features are based on FastText embeddings which encode subword information effectively and can deal with the problem of the OOV words better; graph-based features: We extract the TAGFN as graph-based representations by taking into account user interaction and propagation relationships for improved contextual understanding and classification performance.

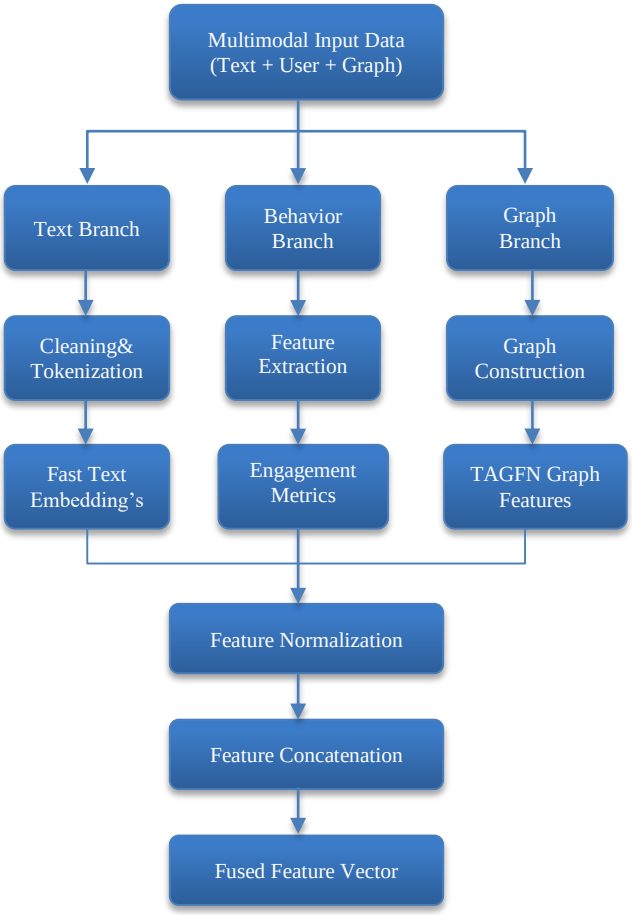


Fig. 2 Multimodal feature extraction and fusion architecture

Structural features, such as node degree, clustering coefficient, and infection depth, are computed in order to characterize news diffusion [32]. This multi-view representation helps the model jointly learn linguistic prompts and diffusion behavior in order to improve the discriminative power of fake and real news. Moreover, normalization of the features is performed to guarantee numerical stability and equalized contribution from different types of features. Lastly, the output features of the written and associated graph are concatenated to represent a feature vector , which is then used as input to an LSTM classifier.

3.4. GWO-Optimized LSTM

An LSTM classifier models temporal dependencies in multimodal fake news data, with its key hyperparameters automatically optimized using the Grey Wolf Optimizer (GWO) to improve performance and reduce manual tuning [33]. The fundamental operations of the LSTM cell are formulated as:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (1)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (2)$$

Here, $f_t$denotes the forget gate, $c_t$the updated cell state, $h_{t-1}$the previous hidden state, and x_tThe current input, while $\sigma(\cdot)$and $\odot$represent the sigmoid activation and element-wise multiplication, respectively. In GWO, each wolf encodes a candidate set of LSTM hyperparameters, and its position is updated as follows:

$$X(t+1) = \frac{X_\alpha + X_\beta + X_\delta}{3} - A \cdot D \quad (3)$$

where X_α, X_β, and $X_\delta$correspond to the three best solutions, and $A$and $D$are coefficient vectors controlling the balance between exploration and exploitation. he fitness of each wolf is evaluated based on the validation accuracy:

$$Fitness = Accuracy_{val}$$

Algorithm 1 outlines the GWO-based hyperparameter optimization process, while the resulting LSTM parameters are summarized in Table 4, and the proposed model architecture is shown in Figure 3.

Algorithm 1. GWO-Based LSTM Hyperparameter Optimization

Input:

Training set D_{train}, validation set D_{val}
Search space bounds $LB, UB$for learning rate, hidden units, dropout, batch size.
Population size N, maximum iterations T

Output:

Optimal hyperparameter vector X^*

1. Initialize a wolf population.
$$X_i = [lr_i, hu_i, dr_i, bs_i], i = 1,2,\dots,N$$
randomly within bounds $LB \le X_i \le UB$.
2. For each wolf X_i, train the LSTM using $D_{train}$and compute fitness:
$$fitness(X_i) = Accuracy(D_{val})$$
3. Identify the best three wolves:
$$X_\alpha, X_\beta, X_\delta$$
4. For iteration $t = 1$to T:
   4.1 Update control parameter:
$$a = 2 - \frac{2t}{T}$$
   4.2 For each wolf $i = 1$to N:
   - Generate random vectors $r_1, r_2 \in [0,1]$
   - Compute coefficient vectors:
$$A = 2ar_1 - a, C = 2r_2$$

- Update position with respect to the three leaders:

$$D_\alpha = | C_1 X_\alpha - X_i |, D_\beta = | C_2 X_\beta - X_i |, D_\delta = | C_3 X_\delta - X_i |$$
$$X_1 = X_\alpha - A_1 D_\alpha, X_2 = X_\beta - A_2 D_\beta, X_3 = X_\delta - A_3 D_\delta$$
$$X_i(t+1) = \frac{X_1 + X_2 + X_3}{3}$$

- Enforce boundary constraints:

$$X_i(t+1) = \min(\max(X_i(t+1), LB), UB)$$

5. Re-evaluate fitness and update. $X_\alpha, X_\beta, X_\delta$.
6. End For
7. Return $X^* = X_\alpha$ as the optimal LSTM hyperparameter set.

Table 4. Optimal LSTM hyperparameters obtained via Grey Wolf Optimization

Parameter	Range	Value
Learning Rate	[0.0001 – 0.01]	0.0017
Hidden Units	[32 – 256]	128
Dropout	[0.1 – 0.5]	0.32
Batch Size	[32 – 128]	64

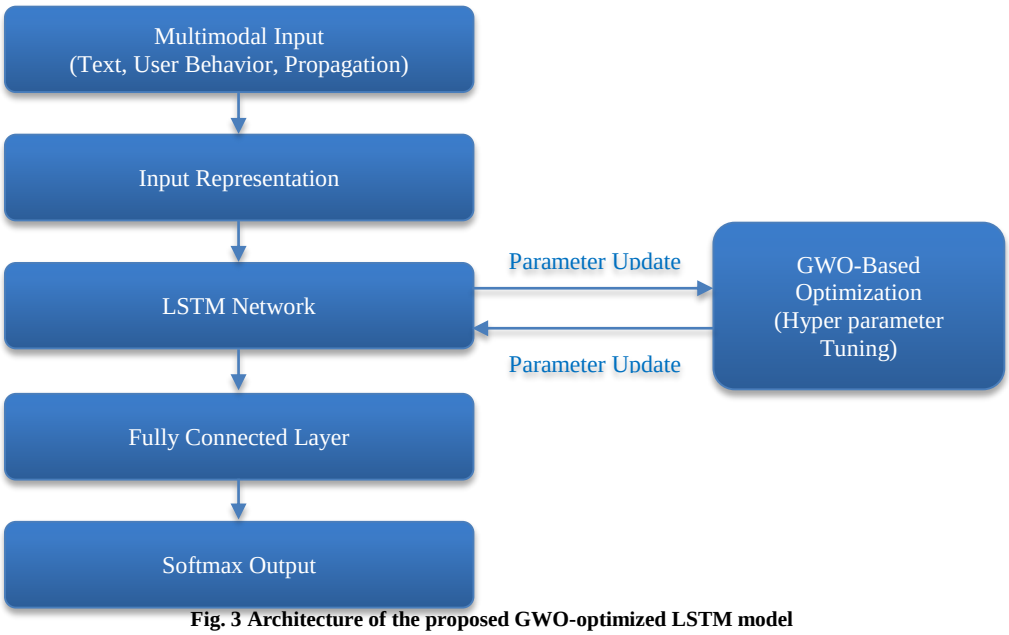


Fig. 3 Architecture of the proposed GWO-optimized LSTM model

3.5. SHAP-Based Explainability

To improve transparency, SHAP is applied to interpret the decisions of the GWO-optimized LSTM model. Instead of viewing the classifier as a black box, SHAP helps reveal how each multimodal feature — including textual embeddings, user-behavior indicators, and TAGFN propagation attributes — influences the final prediction. After the optimized LSTM produces its output, SHAP assigns a contribution score to each feature. These scores indicate whether a specific attribute pushes the prediction toward the fake or real class. This allows us to understand which semantic or structural signals drive the model's decisions.

The explanation can be expressed as:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i x_i \quad (4)$$

Here represents the baseline model prediction, and denotes the contribution of the feature to the prediction. SHAP is used in the current study at the global and instance level: globally, to reveal the most influential multimodal features across valid dataset annotations; locally, to interpret individual classification decisions.

By adopting SHAP in the multimodal model, our model not only gains good performance but also a transparent and reliable decision interpretation [34].

4. Experimental Setup and Performance Evaluation

This section describes the experimental configuration, evaluation criteria, baseline comparisons, and the performance results of the proposed multimodal fake news detection model.

4.1. Experimental Environment

All of our experiments were implemented on a computer with an Intel Core i7, 32GB of RAM, and an NVIDIA RTX 3080 GPU. The model was implemented in Python using the TensorFlow and Keras libraries. Textual representation of

each feature was embedded using FastText, and structural features were extracted from the TAGFN graph using NetworkX. Interpretability was analyzed by SHAP.

4.2. Evaluation Metrics

We evaluated model performance according to standard classification metrics (Equation (5)–(7). Robustness evaluation under adversarial perturbations through computing a Robustness index (RI): RI as defined in Equation (8), where ASR indicates the attack success rate [35].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

To assess robustness, the Robustness Index (RI) was calculated as:

$$RI = 1 - ASR \quad (8)$$

where ASR is the attack success rate.

4.3. Baseline Models

The GWO-optimized LSTM model we proposed was compared with other baseline models, including TF-IDF + SVM, FastText + LSTM, CNN–BERT, and GCN-based architecture. For our experiment, we used both Multi-Fake-DetectiVE 2023 and TAGFN datasets. Results summary. Here we summarize the results above in Tables 5 and 6.

As shown in Table 5, the comparative results evidence how much the overall prediction accuracy and balance of precision and recall for the Multi-Fake-DetectiVE 2023 dataset that the proposed method enjoys compared to baselines.

Table 5. Performance on Multi-Fake-Detection 2023

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TF-IDF + SVM	90.34	89.82	88.9	89.35
FastText + LSTM	94.91	94.3	94.05	94.17
CNN-BERT	96.88	96.42	96.01	96.21
GCN	95.97	95.31	95.1	95.2
Proposed	98.71	98.63	98.52	98.57

Similar trends can be observed in Table 6 on the TAGFN dataset as well, where multimodal architecture consistently improves across any performance metric, further establishing its robustness and generalization ability over heterogeneous distribution of social media data.

Table 6. Performance on TAGFN Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TF-IDF + SVM	91.69	90.4	90.02	90.21
FastText + LSTM	95.86	95.35	95.2	95.27
CNN-BERT	97.54	97.1	96.92	97.01
GCN	96.73	96.14	95.88	96.01
Proposed	99.32	99.48	99.36	99.42

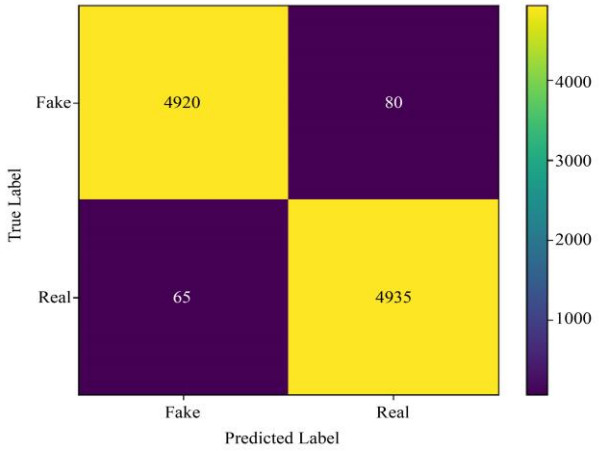
4.4. Confusion Matrix and ROC Analysis

Figure 4 displays the confusion matrices on Multi-Fake-DetectiVE 2023 and TAGFN datasets. We can see from the results that most of the predictions are concentrated around the main diagonal, indicating right classification for fake and real news samples.

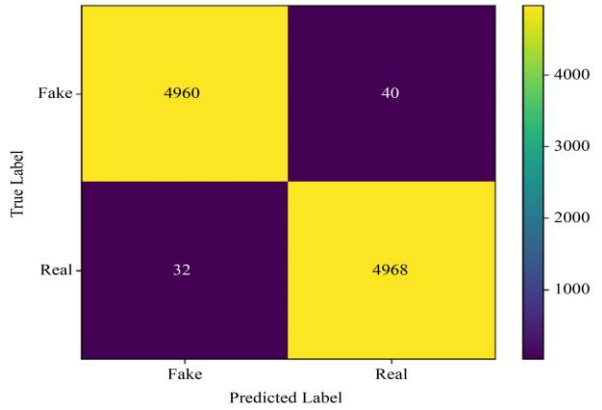
The diagonal dominance here suggests that the model is clearly separating both classes with high confidence.

The ROC curves generated are shown in Figure 5, which exhibited stable performance across datasets. So the curves imply a high true positive rate and low false positive rate for given thresholds, which indicates good discriminative power. Based on the heatmap, we have very few off-diagonal elements, which indicates that our misclassification rates are low. This suggests that the model does well on both classes and is getting good precision, recall, and general accuracy scores.

In addition, solid performance in both the confusion matrices and ROC curves results from using multimodal features as well as the optimization process. Textual, behavioral, and propagation-based information further collectively help the model identify complex patterns, while GWO allows self-optimized parameters and provides stable convergence. It leads to better decision boundaries and reduced variance of class labels across different data distributions.

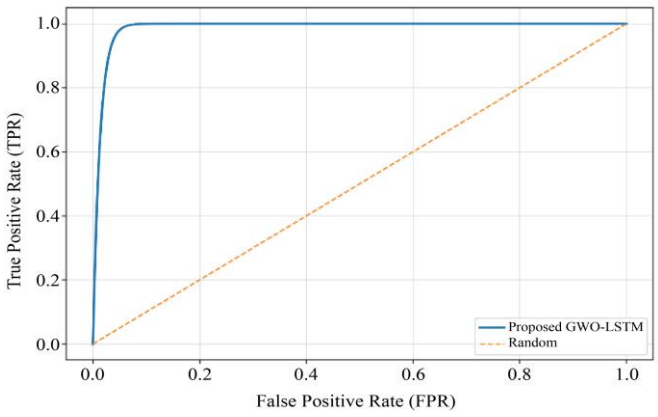


(a)

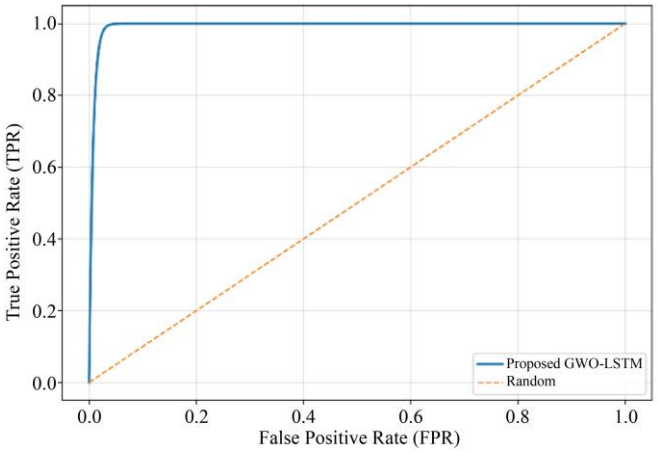


(b)

Fig. 4 Confusion matrices of the proposed GWO-optimized LSTM model on (a) Multi-Fake-DetectiVE 2023 and (b) TAGFN datasets



(a)



(b)

Fig. 5 ROC curves of the proposed GWO-optimized LSTM model on (a) Multi-Fake-DetectiVE 2023 and (b) TAGFN datasets

4.5. Class-Wise Evaluation

Results for class-wise performance on Multi-Fake-DetectiVE 2023 and TAGFN are shown in Table 7. Results: The table displays consistently high precision, recall, and F1-score values across fake and real classes.

The reason for balanced performance is that the model is not biased towards either class because of stratified sampling and multimodal feature integration. We combine textual, behavioral, and Propagation features so that we can discriminate the classes effectively.

The fact that the model has such high precision and recall values means it is able to avoid making false predictions as much as possible while still being capable of accurately marking true instances. As a result, GWO-based optimization further enhances the overall performance by improving parameter tuning and generalization.

Table 7. Class-wise Performance on Both Datasets

Dataset	Class	Precision	Recall	F1-score
Multi-Fake-DetectiVE 2023	Fake	0.99	0.992	0.991
Multi-Fake-DetectiVE 2023	Real	0.989	0.988	0.988
TAGFN	Fake	0.991	0.993	0.992
TAGFN	Real	0.99	0.989	0.989

4.6. Model Validation

In this section, we analyze (i) the contribution of feature groups, (ii) the effect of GWO-based tuning, and (iii) robustness against adversarial attacks as shown in the ablation study (Table 8). TAGFN + user-behavior+textual features can achieve the best performance. The findings demonstrate that task performance increases with the inclusion of additional feature groups, confirming that both semantic and structural signals are complementary. It shows that text-based methods, processes, and approaches alone for fake news detection are insufficient at an adversarial level.

In Table 9, we compare the model performance prior to GWO-based hyperparameter optimization and post-hyperparameter optimization in order to show how optimal parameter selection impacts the model performance.

Table 8. Ablation Study Results for Different Feature Combinations

Configuration	Accuracy
Text Only	0.954
Text + User Features	0.971
Text + TAGFN	0.982
Full Model	0.99

Table 9. Effect of GWO Optimization

Model	Accuracy Before GWO	Accuracy After GWO
LSTM	0.964	0.99

Robustness against adversarial attacks is assessed under different adversarial threat models. Stability results over all attack types appear in Table 10. The attack success rate stays within a limited range, and the robustness index achieves nearly the same values, indicating stable performance against attracted engagement and semantic perturbations.

Table 10. Robustness Under Adversarial Attacks

Attack Type	ASR	RI
Bot Injection	0.058	0.942
Semantic Perturbation	0.071	0.929
Network Manipulation	0.063	0.937

4.7. Comparison with State-of-the-Art

Comparison with literature in Table 11 represents the comparison of the proposed model versus several recent state-of-the-art methods. In this regard, Study [36] proposes a unified self-learning multimodal framework that combines contrastive learning from a large language model to learn inter-modality features. While this technique improves the state of the art in multimodal representation and reduces reliance on labeled data, it only focuses on explicit content-level fusion without incorporating user behavior or propagation dynamics [22]. In [37], a Hybrid CNN–BERT Model with Grey Wolf Optimization (GWO) is proposed to improve feature extraction and classification.

This approach takes advantage of optimized parameter tuning with improved semantic representations, but is still limited to text features and doesn't capture structural or behavioral data. Graph Neural Networks are applied for modeling structural relationships in news propagation networks [38]. This method effectively exploits graph-based relationships and achieves high detection performance, but is limited to structural patterns without the incorporation of semantic and multimodal features. Study [39] presents a propagation dynamics-based framework that simulates information diffusion to extract dynamic features from large language models.

While this method increases the understanding of propagation behavior, it operates with simulated dynamics and does not encompass actual interaction characteristics in a single multimodal model. Combining interpretable machine learning methods with SHAP-based Explainability in [40]. Higher transparent option, but zero detection, and the modeling of hard multi-modal relationships.

However, the proposed model challenges this notion and uses semantic embeddings UK-bIs features representing user-behavior indicators and TAGFN propagation to complement the branch of GWO-optimized LSTM architecture. By optimizing multimodal signals with evolutionary hyperparameter tuning, generalization and convergence are improved. The classification accuracy of this design over these test datasets is 99.32% shown in Table 11.

Table 11. Comparison with State-of-the-Art Methods

Method	Dataset	Feature Type	Model	Performance
Self-Learning Multimodal [36]	Public multimodal dataset	Text + Image (Multimodal)	LLM + Contrastive Learning	Acc/Prec/Rec/F1: >85%
BERT-GWO-CNN [37]	Social media datasets	Text (Semantic Features)	BERT + CNN + GWO	High accuracy (outperforms RF, SVM, RNN, LSTM)
GNN-Based Model [38]	Social media propagation networks	Graph-based (Structural)	GNN + Deep Learning	Acc: 97%
ProFNSE [39]	Chinese–English datasets	Text + Simulated Propagation	XGB / LR Classifiers	Acc: 0.888 (CN), 0.9717 (EN)
SHAP-XAI Model [40]	Benchmark datasets	Text + Explainability	ML + SHAP	Improved interpretability

4.8. Discussion

The findings indicate the importance of aggregating textual, behavioral, and Propagation information for fake news detection tasks. Models using pure text have some success but find the complexity of misinformation very constrained. The use of user interaction features, as well as transitive propagation structures such as TAGFN, provides complementary information that aids in better performance than using only linguistic content. This intuition is strengthened by the fact that linguistic signals should be jointly examined alongside diffusion behavior to improve our understanding of misinformation events.

The ablation analysis shows that each of the feature groups contains valuable patterns and finds that propagation and engagement indicators are complementary sources to semantic embeddings. In more detail, these heterogeneous features are merged to learn more discriminative and meaningful representations compared with those based solely on text-centric approaches.

The application of GWO for automatic optimization stabilizes convergence and generalization relative to models in which hyperparameter tuning is applied manually. This approach minimizes the risk of poor parameter choice and increases performance across models.

Thus, the robustness test strengthens the evidence about the multimodal framework's robustness, indicating that compared to the normal methods, manipulation strategies have a relatively less effect on multimodal frameworks under adversarial attacks. Moreover, including SHAP-based Explainability improves the transparency of the model through understanding feature contributions. These results confirm that predictions are most sensitive to propagation depth and interaction-related features, which reinforce conclusions about the advantage of combining structural and behavioral information with semantic features.

That is already a strength of this evaluation, but it may be limited to English-language datasets, and the computational requirements could be in the way of a mass-real-time deployment. Hence, future work should aim at broadening the scope of multilingual datasets and increasing computational efficiency to handle practical applications in real-world scenarios.

4.9. Explainability Analysis

Figure 6 shows the SHAP summary plot for the best-performed multimodal model. The findings reveal that propagation depth and bot-related engagement features have the greatest impact on categorization, followed by key semantic embedding dimensions based on syntactic representations. This distribution indicates that structural diffusion signals indeed carry important discriminative power for discerning between fake news and true news.

At an instance level, the SHAP values show how particular feature combinations push predictions towards the fake or real class. Samples that have abnormal patterns of Propagation and samples with high intensity of bot interactions always make a positive contribution to the fake label. These results are consistent with the behavioral traits that have been associated with the spreading of misinformation. The explainability results confirm the consistency of the multimodal design, and provide evidence for embedding semantic and Propagation features in the optimized LSTM model.

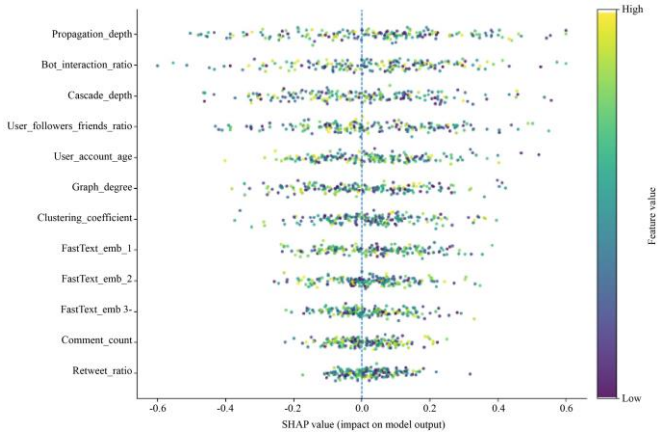


Fig. 6 SHAP summary plot of feature contributions in fake news detection

5. Conclusion

This study presented a multimodal fake news detection framework that integrates textual semantics, user behavioral patterns, and propagation dynamics within a Grey Wolf Optimizer–enhanced LSTM architecture supported by SHAP-based interpretability. By combining optimization strategies with explainable modeling, the proposed framework effectively overcomes the limitations associated with single-modality approaches and black-box deep learning models.

Experimental evaluation on the Multi-Fake-Detection 2023 and TAGFN datasets demonstrates that the proposed model achieves superior and consistent performance compared to well-established baseline methods across multiple evaluation metrics. The model achieves a classification accuracy of 99.32% and is resilient to adversarial settings, suggesting the reliability in capturing an intricate pattern of misinformation.

The ablation analysis enhances the effect of multimodal integration around feature fusion, but the explainability module highlights that not only are first-order propagation properties important features that influence decision outcome, but so also are users and semantic indicators at higher orders. Second, multiple runs statistical validation also gives confidence about the stability of the proposed framework with respect to random initialization tendency for different performances. It also demonstrates strong transfer learning

performance across different datasets, suggesting the model is generalizable to other social media contexts. We also show through computational analysis that the optimization can converge rapidly with modest compute cost, lending a practicality to the framework for real-world deployment scenarios.

5.1. Limitations and Future Work

The proposed framework works under ideal conditions, but many factors add layers of constraints. Diversity of real-world misinformation: Your experimental evaluation only includes a handful of benchmark datasets, failing to show that your approach is general enough to effectively handle the many variations of misinformation created across different domains. Thus, domain generalization of the model and testing on new age social platforms where we have never trained information might be necessary.

Second of all, the additional joint multimodal features do help in improved detection, but at the same time introduce large computational complexities that affect the system scalability and real-time conversion on large-scale computation systems. In addition, existing systems trained on English-language datasets also need manual efforts to generalize in cross-lingual and multicultural settings.

Those who trained on it are limited by its reliance on propagation-based features, not necessarily present or consistent within early-stage misinformation detection. These information streams can change rapidly, and incorrect partial/noisy propagation data can affect the model performance.

In more detail, future work will aim to apply our high-level framework to multilingual datasets and include transformer-based architectures shown to generate better contextual representation learning. Future work will focus on how to improve the computational efficiency of the model to enable deployment in real-time in social media settings. The released version of the system should also provide further research into adversarial robustness that makes it robust to strategies employed by malicious actors, misinformation campaigns, or coordinated manipulation attacks.

References

[1] Faisal A. Alshuwaier, and Fawaz A. Alsulaiman, "Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives," *Computers*, vol. 14, no. 9, pp. 1-40, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Pingping Yang et al., "Multi-Modal Transformer for Fake News Detection," *Mathematical Biosciences and Engineering*, vol. 20, no. 8, pp. 14699-14717, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo, "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection," *Information*, vol. 13, no. 12, pp. 1-28, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Carmela Comito, Luciano Caroprese, and Ester Zumpano, "Multimodal Fake News Detection on Social Media: A Survey of Deep Learning Techniques," *Social Network Analysis and Mining*, vol. 13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alim Al Ayub Ahmed et al., "Detecting Fake News Using Machine Learning: A Systematic Literature Review," *arXiv preprint*, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] G. Bharath et al., "Detecting Fake News Using Machine Learning Algorithms," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Hounaida Moalla et al., "Exploring the Power of Dual Deep Learning for Fake News and AI-Generated News Detection," *Informatica*, vol. 48, no. 4, pp. 567-594, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Amila Silva et al., "Unsupervised Domain-agnostic Fake News Detection Using Multi-modal Weak Signals," *arXiv preprint*, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Jiaying Wu, Jiafeng Guo, and Bryan Hooi, "Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks," *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona Spain, pp. 3367-3378, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] M. F. Mridha et al., "A Comprehensive Review on Fake News Detection with Deep Learning," *IEEE Access*, vol. 9, pp. 156151-156170, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Changhe Song et al., "CED: Credible Early Detection of Social Media Rumors," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3035-3047, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Arun Kumar Yadav et al., "Fake News Detection using Hybrid Deep Learning Method," *SN Computer Science*, vol. 4, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yexin Tian et al., "An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection," *Mathematics*, vol. 13, no. 13, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Khurram Shahzad et al., "A Scoping Review of the Relationship of Big Data Analytics with Context-Based Fake News Detection on Digital Media in Data Age," *Sustainability*, vol. 14, no. 21, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse, "A Survey of Explainable AI Techniques for Detection of Fake News and Hate Speech on Social Media Platforms," *Journal of Computational Social Science*, vol. 7, pp. 587-623, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ajay Kumar, and James W. Taylor, "Feature Importance in the Age of Explainable AI: Case Study of Detecting Fake News & Misinformation via a Multi-modal Framework," *European Journal of Operational Research*, vol. 317, no. 2, pp. 401-413, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ziyi Zhou et al., "FineFake: A knowledge-enriched Dataset for Fine-grained Multi-domain Fake News Detection," *Information Fusion*, vol. 132, pp. 1-12, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[18] Lidong Wang et al., "Multimodal Fusion with LLM Content via Hierarchical Progressive Transformer for Explainable Fake News Detection," *Information Processing & Management*, vol. 63, no. 5, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[19] Pervaiz Akhtar et al., "Detecting Fake News and Disinformation using Artificial Intelligence and Machine Learning to Avoid Supply Chain Disruptions," *Annals of Operations Research*, vol. 327, pp. 633-657, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] A. Kumari Shalini, Sameer Saxena, and B. Suresh Kumar, "Designing A Model for Fake News Detection in Social Media Using Machine Learning Techniques," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2s, pp. 218-226, 2023. [Google Scholar] [Publisher Link]

[21] Ciprian-Octavian Truic, and Elena-Simona Apostol, "It's all in the Embedding! Fake News Detection Using Document Embeddings," *Mathematics*, vol. 11, no. 3, pp. 1-29, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth, "Robust Fake News Detection using Large Language Models under Adversarial Sentiment Attacks," *Proceedings of the ACM Web Conference*, Dubai United Arab Emirates, pp. 1717-1726, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[23] Pawan Kumar Verma et al., "MCred: Multi-modal Message Credibility for Fake News Detection using BERT and CNN," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 10617-10629, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Zhiwei Guo et al., "A Novel Fake News Detection Model for Context of Mixed Languages Through Multiscale Transformer," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5079-5089, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jing He et al., "FACTGUARD: Event-Centric and Commonsense-Guided Fake News Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 1, pp. 363-371, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[26] Shankar Biradar, Sunil Saumya, and Arun Chauhan, "Combating the Infodemic: COVID-19 Induced Fake News Recognition in Social Media Networks," *Complex & Intelligent Systems*, vol. 9, pp. 2879-2891, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Junfei Wu et al., "Adversarial Contrastive Learning for Evidence-Aware Fake News Detection with Graph Neural Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5591-5604, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Lixia Ji et al., "A Local Dynamic Propagation Graph-based Method for Multi-modal Fake News Detection," *Journal of Intelligent Information Systems*, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[29] Alessandro Bondielli et al., "MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task," *Proceedings 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023)*, Parma, Italy, vol. 3473, pp. 1-8, 2023. [Google Scholar] [Publisher Link]

[30] Kay Liu et al., "TAGFN: A Text-Attributed Graph Dataset for Fake News Detection in the Age of LLMs," *arXiv preprint*, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[31] Siyu Lu et al., "Multiscale Feature Extraction and Fusion of Image and Text in VQA," *International Journal of Computational Intelligence Systems*, vol. 16, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Delvin Ce Zhang, et al., "Text-Attributed Graph Representation Learning: Methods, Applications, and Challenges," *Companion Proceedings of the ACM Web Conference 2024*, Singapore, pp. 1298-1301, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Saif Alsudani et al., "Enhancing Spam Detection: A Crow-Optimized FFNN with LSTM for Email Security," *Wasit Journal of Computer and Mathematics Science*, vol. 3, no. 1, pp. 28-39, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Prokash Gogoi, and J. Arul Valan, "Enhancing Date Fruit Classification using Machine Learning, CTGAN, and SHAP-based Explainability," *Journal of Food Measurement and Characterization*, vol. 19, pp. 6851-6872, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[35] Saif Wali Ali Alsudani, and Ghassan Khudair Saud, "Recurrent Neural Network Optimized by Grasshopper for Accurate Audio Data-Based Diagnosis of Parkinson's Disease," *Wasit Journal for Pure Sciences*, vol. 4, no. 2, pp. 56-75, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[36] Hao Chen et al., "A Self-learning Multimodal Approach for Fake News Detection," *Frontiers in Artificial Intelligence*, vol. 8, pp. 1-25, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[37] Rajalakshmi Viswanathan et al., "Detection of Fake News in Social Media using CNN with Grey Wolf Optimized BERT," *Acta Scientiarum. Technology*, vol. 47, pp. 1-11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[38] Haji Gul et al., "Advancing Fake News Detection with Graph Neural Network and Deep Learning," *Journal of Physics: Complexity*, vol. 6, no. 2, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[39] Fuqiang You et al., "ProFNSE: Propagation Dynamics-derived Fake News Detection in Social Networks," *The Journal of Supercomputing*, vol. 81, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Sanya Bansal, Harshit Panwar, and Geetika Munjal, "Detecting Fake News with XAI: A Look at SHAP for Making Machine Learning Models More Explainable," *2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON)*, Pune, India, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]