scientific reports

Check for updates

OPEN

A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization

Ramesh Kumar Ayyasamy[1]✉, Chinnasamy Ponnusamy[2], Kovvuri N. Bhargavi[3], Srikanth Cherukuvada[4], G. Charles Babu[5], S. Amutha[6] & Dawit Tadesse Gamu[7]✉

In recent years, the widespread dissemination of fake news on social media has raised concerns about its impact on public opinion, trust, and decision-making. Addressing the limitations of traditional detection methods, this study introduces a hybrid deep learning approach that enhances the identification of fake news. The objective is to improve detection accuracy and model robustness by combining a Long Short-Term Memory (LSTM) network for contextual feature extraction with a Convolutional Gaussian Perceptron Neural Network (CGPNN) for classification. To further optimize performance, we integrated a metaheuristic Moth-Flame Whale Optimization (MFWO) algorithm for hyperparameter tuning. Experimental evaluation was conducted on four benchmark datasets ISOT, Fakeddit, BuzzFeedNews, and FakeNewsNet using standardized preprocessing techniques and TF-IDF-based text representation. Results show that the proposed model outperforms existing methods, achieving up to 98% accuracy, 95% F1-score, and statistically significant improvements ($p < 0.05$) over transformer-based and graph neural network models. These findings suggest that the hybrid framework effectively captures linguistic patterns and textual irregularities in deceptive content. The proposed method offers a scalable and efficient solution for fake news detection with practical applications in social media monitoring, digital journalism, and public awareness campaigns. Overall, the framework delivers 3–8% higher accuracy and F1-score compared to state-of-the-art approaches, demonstrating both robustness and practical applicability for large-scale fake news detection.

Keywords Social media, Fake news detection, Feature extraction, Deep learning, Convolutional gaussian perceptron neural network, Metaheuristic optimization, LSTM, Classification

Fake news has existed long before the advent of digital technology, with the deliberate dissemination of false information dating back to ancient times. However, the proliferation of internet technologies and computational advancements has dramatically transformed the landscape of information sharing. Contemporary digital platforms particularly social media networks have created unprecedented opportunities for content generation and dissemination with minimal barriers to entry[1]. The information revolution has brought about democratization of information access, yet it has also enabled fast dissemination of both genuine and false content. The replacement of traditional media channels with social media as primary information sources has led to the fast spread of misleading content. False information spreads past its original targets to affect society at large while damaging public trust in authentic news sources and creating false public reactions to factual reporting. Research showed that fabricated content spread more widely on Facebook and Twitter than accurate reporting during

[1]Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia. [2]Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Srivilliputtur, India. [3]Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India. [4]Department of CSE (AI & ML), B V Raju Institute of Technology, Narsapur, India. [5]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, Telangana, India. [6]Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India. [7]Department of Computer Science, School of Computing and Informatics, Mizan Tepi University, Mizan Teferi, Ethiopia. ✉email: rameshkumar@utar.edu.my; dawittadesse@mtu.edu.et

the 2016 U.S. presidential election[2]. The September 2024 Springfield pet-eating hoax spread false information about Haitian immigrants eating domestic pets after political figures shared it despite its baseless origins from an unsubstantiated social media post[3]. The financial incentives behind fake news proliferation should not be ignored. Research shows that major technology platforms gain indirect advantages from users engaging with provocative false content. Websites that produce fabricated news generate significant revenue through online advertising systems which creates financial incentives for spreading misinformation[4]. The financial aspect became clear in July 2024 when false information about a Southport tragedy led to civil disturbances throughout the United Kingdom[5]. Unsubstantiated claims about government fund misappropriation to media outlets which independent fact-checkers later disproved demonstrate how misinformation affects public discourse at its highest levels[6].

Specialized fact-checking websites and platform-integrated tools such as the "community notes" system implemented on X (formerly Twitter) have emerged to fight this trend. The International Federation of Library Associations and Institutions has created frameworks to help users detect unreliable content and Bozkurt et al.[8] have conducted systematic assessments of current detection and prevention methods. These initiatives recognize that political strategies frequently build upon misinformation which affects financial markets and investment choices and crisis management. The intentional creation of fake news that looks authentic creates major obstacles for detection systems[9]. Social media platforms enable users to share content with their connected network which leads to increased potential impact[10]. The 2016 survey showed that 62% of American adults obtained news from social media platforms while 49% did in 2012 and 47% used social media as their main news source thus making fake news exposure both widespread and dangerous[10]. The situation demands immediate implementation of effective protective measures against misinformation. The implementation of complete detection systems faces multiple technical obstacles because of reference datasets and event coverage and consumption patterns and verification processes and content divergence[11]. The research community has developed multiple solutions to tackle these issues yet fake news detection accuracy remains a persistent challenge which drives ongoing investigations into better detection methods. The rapid growth of false information online requires immediate development of automated systems which can analyze large volumes of content. Deep learning techniques show exceptional potential when used for social network content evaluation[12].

Traditional fake news detection methods have mainly depended on content analysis of news articles' intrinsic features while social context models that study information diffusion patterns have been adopted in recent times[13]. The enormous amount of content and its fast spread across platforms makes manual assessment impossible so automated systems must be developed to quickly assess information reliability. The development of automated models has focused on either news content or social context features. The approaches use data mining algorithms to extract fake news characteristics which are based on established social and psychological theories. A general classification model for fake news identification consists of two stages: feature extraction and model construction from a data mining perspective. The system extracts relevant content characteristics during feature extraction before using these representations to differentiate between authentic and fabricated news in the model construction phase[14,15,47–49].

Major contribution of the work

This research makes several significant contributions to the field of fake news detection:

- A new hybrid method will be proposed which combines deep learning models with metaheuristic algorithms for feature extraction and classification that is specifically designed for social media fake news identification.
- We establish a systematic feature selection process to determine the most discriminative indicators of fake news content.
- We use recurrent LSTM networks to extract temporal and contextual features from textual data to capture the subtle patterns that are characteristic of fabricated content. We develop a convolutional Gaussian perceptron neural network (CGPNN) architecture for classification, combining the strengths of convolutional feature extraction with probabilistic modeling.
- The MFWO algorithm optimizes model performance by improving both parameter selection and classification results.
- ISOT Dataset → Achieved 92% accuracy, around 3% higher than transformer-based baselines and 10–14% higher than traditional models (CNN, DT-RF).
- Fakeddit Dataset → Reached 95% accuracy and 93.5% F1-score, surpassing existing deep learning approaches by 4–6%.
- BuzzFeedNews Dataset → Delivered the best performance with 98% accuracy and 95% F1-score, improving on advanced GNN and multimodal models by 2–5%.
- FakeNewsNet Dataset → Attained 96.5% accuracy, 94.8% precision, and 95.5% recall, showing consistent improvements across evaluation metrics.
- Overall → The proposed hybrid LSTM–CGPNN with MFWO optimization outperforms state-of-the-art methods by 3–8% on average, with statistically significant gains ($p < 0.05$).

These outcomes demonstrate the model's effectiveness in capturing deceptive patterns and underline its practical potential for real-world fake news detection.

Literature review

Current approaches in fake news detection

Detection of misinformation has become a key research domain, and several different approaches have been suggested that leverage diverse methods and data sources. Li et al.[16] proposed an MH (Machine-Human)

framework that integrates machine learning and network-based methods with metrics specifically designed for human literacy, thereby developing an end-to-end framework for the detection of misinformation on social media platforms. Their approach showed how computer-based methods augmented with human-driven assessment protocols can build detection capabilities more than the sum of what is feasible for either technique standalone. Azam et al.[17] utilized Natural Language Processing (NLP) methodologies coupled with Bidirectional Encoder Representations from Transformers (BERT) and human-in-the-loop coding for detecting misinformation in various languages. Through the application of their models to datasets with English, Arabic, and Urdu content, they demonstrated the utility of transformer-based models for cross-lingual fake news detection. Their results highlight the necessity of language-specific considerations for universal detection system design. Khanday et al.[18] made a comparative study of several neural network architectures, namely Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Bidirectional LSTM (Bi-LSTM), Character-level LSTM (C-LSTM), Hierarchical Attention Networks (HAN), and the Convolutional Hierarchical Attention Network (Conv-HAN) to compare the efficiency of various embedding techniques. They compared Global Vectors (GloVe) with character embeddings on several datasets and determined the most suitable representations for a range of content types and linguistic models. While analyzing ensemble methods, Comito et al.[19] applied a hybrid of Deep Learning and ensemble methods to identify true and fake news on various domains. Through their comparative study, it was observed that the XGBoost ensemble method outperformed single classifiers, indicating the benefits of aggregating multiple models for enhancing detection accuracy.

Advanced models and architectures

Ruchansky et al.[20] suggested the CSI (Capture, Score, and Integrate) model for fake news automatic detection. The hybrid model utilizes Recurrent Neural Networks (RNN) to capture individual behavior (users and articles) and collective user interactions involved in the propagation of fake information. Content and behavioral patterns are both integrated by the CSI approach, which provides a deeper analysis than content-based methods. In the field of multimodal detection, Albalawi et at[21]. proposed a two-stream deep learning architecture called Coupled ConvNet. The framework processes both visual and textual data using separate CNN streams before fusing the results using weighted combination for evaluation. Their approach addresses the increasing phenomenon of deceptive text being complemented by misleading images in fake news dissemination. Several researchers have investigated GCN-based approaches to rumor detection and misinformation propagation on social media[22–26]. These approaches represent the diffusion patterns of information as graphical representations, and in doing so, facilitate the examination of interrelations among content pieces, users, and interaction patterns. Various datasets derived from Twitter and Weibo have been utilized to validate the utility of graph-based models in identifying the structural patterns that are indicative of fake news propagation. Do et al.[27] proposed a multi-entry neural network architecture (MENET) to model different kinds of data that aimed to predict the geographical location of Twitter users. Although effective, Graph Convolutional Neural Networks have increased computational demands both temporally and spatially[54]. To alleviate this issue, Xu et al.[28] proposed an improved model of Graph Neural Networks called Hierarchically Aggregated GNN (HAGNN) for rumor detection with accuracies of 95.7% and 88.2% on the Weibo dataset.For feature extraction, Sharma et al.[25] suggested the Fake News Ratio as a novel measure aimed at extracting unique features from real-world datasets. Their method came with extra complexities that had to be improved.

Hybrid and context-based approaches

Dixit et al.[29] suggested a novel approach to the categorization of social media misinformation based on the application of the Recurrent Convolutional Neural Network (RCNN) framework. The approach successfully captures contextual semantic details of messages and subsequently learns sentiment representations via its recurrent model, enabling further analysis of potentially deceptive content.

To leverage semantic discrepancies between news article titles and their content a common characteristic of fake news Gorai & Shaw[30] developed an innovative feature selection technique that identifies multiple semantic dissimilarity patterns. By quantifying these inconsistencies as distance values and incorporating them as classification features, their approach targets a key indicator of fabricated content. However, the computational requirements of RCNN architectures present challenges for real-time detection applications.

Ramya et al.[31] proposed an attention-dependent bidirectional Gated Recurrent Units (GRU) model to extract attributes from news content alongside a deep framework for extracting latent representations from auxiliary data. Their system then learns an attention distribution across these hidden vectors, concatenating them into an attention matrix for classification.

Chen et al.[32] surveyed various veracity assessment techniques for online fake news detection, exploring two primary assessment categories: methods based on linguistic cues and approaches utilizing network analysis. Their review included classifier training using Support Vector Machines (SVM) and naïve Bayes models with linguistic features, as well as network-based methods incorporating linked data processing and social behavior analysis.

Zhong et al.[33] used a novel approach to investigate the behavior of news propagation on social media websites, and more specifically the dynamics of rumor propagation on Twitter and discovering unique propagation patterns that discriminate between fake news and real journalism.

Jadhav and Singh[14] investigated means of identifying fake news by analyzing users' comments, utilizing bipartite network experiments with the support of user feedback to identify patterns suggesting misinformation. Likewise, Alshahrani et al.[34] proposed a framework for verifying the truth of events tweeted on Twitter, offering comparative evaluations of current methodologies and their proposed method.

Recent advances in multimodal and optimization-based detection

Several researchers have investigated hybrid approaches that integrate various sources of information[51–53]. Another significant contribution in this direction is the one by Zhang et al.[13], who presented a DeepNet-based fake news detection method, tested against actual BuzzFeed and Fakeddit datasets. The approach incorporates tensor factorization for fusing social context information with news content features, demonstrating that such an integrated approach provides improved detection accuracy over content- or context-only methods.

The usage of metaheuristic optimization algorithms in the domain of fake news detection is a novel direction in this research. Yildirim[35] used two metaheuristic algorithms, Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO), for analyzing the content of fake news and obtained 87.14% and 71.61% accuracy, respectively, on a large-scale news dataset. These findings are indicative of the prospects of optimization-based methods in improving detection efficacy.

Nadeem et al.[36] demonstrated a multimodal network model proficient in executing diverse levels and kinds of data fusion, incorporating news title text, metadata, and additional related information from the Fakeddit corpus. Their research emphasizes the value of taking multiple content modalities into account when developing effective detection systems.

Liang et al.[37] developed a machine learning model that incorporates user characteristics, news content, and social network dynamics grounded in the concept of social capital. Using the XGBoost algorithm to evaluate feature importance, they identified key factors influencing fake news detection and implemented comparative analyses of several classification models SVM, Random Forest, Logistic Regression, CART, and Neural Networks to determine optimal approaches.

Guo et al.[38] proposed a method for identifying rumor sources using limited observations by calculating rumor severity, dissemination centrality, and diffusion propensity at network nodes. By employing Gaussian density distributions and infection distance metrics to assess the likelihood of each node being a source, their approach achieved strong correlations with the accuracy rates of various classification models, including Random Forest (73.9%), Naïve Bayes (80.8%), and Passive Aggressive (81%).

Deep learning in fake news detection: challenges and opportunities

The advent of social media websites has massively increased the spread of misinformation, thereby necessitating the creation of more advanced detection tools using deep learning (DL) techniques. Despite the proven effectiveness of DL techniques such as CNNs, LSTMs, and BERT in detecting complex linguistic patterns, many existing models are trained on domain-specific, mostly English-language datasets such as ISOT and FakeNewsNet, thereby limiting their generalization potential across diverse cultural or subject matter domains[7]. Transcending these limitations, our research suggests a Deep Learning ensemble approach founded on the Convolutional Gaussian Perceptron Neural Network (CGPNN), which connects CNN's locality of features with probabilistic reasoning for improved detection of latent patterns within deceptive text. The suggested framework aims to exceed the limitations pertaining to static forms and strengthen classification using ensemble diversity, an initiative demonstrated to be more effective compared to traditional unimodal approaches in recent investigations[53]. Moreover, grounded on multimodal detection methods as exemplified by MFFFND-Co through semantic alignment of textual and visual information, our model is suggested for future pursuit in cross-modal fields, thereby promoting versatility in real-world applications wherein misinformation is becoming common in hybrid media[54]. Lastly, the ensemble-based CGPNN approach offers a scalable, explainable, and cross-domain-agnostic remedy to the existing limitations in fake news detection.

Research gaps identified

- *Infrequent Integration of Deep Learning and Metaheuristic Optimization*: Although deep learning models like CNNs, LSTMs, and transformers are widely used, there is a scarcity of research that combines them with metaheuristic optimization algorithms. Previous attempts to use optimizers such as PSO and GWO have generally achieved only moderate success and have not been fully integrated into sophisticated deep learning frameworks.
- *Limited Exploration of Probabilistic and Distributional Models*: Current models, including graph-based and transformer networks, excel at capturing contextual and relational data. However, probabilistic methods, such as those based on Gaussian distributions, have been largely overlooked. These approaches could offer a way to model deceptive text's inherent uncertainty and unique statistical properties.
- *Absence of a Unified Framework for Content and Context*: Research in fake news detection is often divided into two distinct streams: content-based features (like linguistic patterns) and context-based features (such as social network propagation). A comprehensive, holistic model that effectively integrates content and context analysis remains underdeveloped.
- *Practical Challenges in Multimodal Detection*: While some studies have begun to fuse text and images to detect fake news, these multimodal systems frequently suffer from high computational complexity. This complexity creates scalability issues, making them difficult to implement for the real-time detection required on social media platforms.
- *Constraints of Cross-Lingual and Cross-Domain Applicability*: Most standard benchmark datasets, including ISOT, FakeNewsNet, BuzzFeedNews, and Fakeddit, are heavily biased towards English-language, political news. This narrow focus severely restricts the generalizability of trained models, limiting their effectiveness in other languages, cultural settings, and subject areas.
- *Vulnerability to Sophisticated and Hybrid Misinformation*: Existing detection systems are not robust enough to handle nuanced forms of false content. They can be easily bypassed by satirical articles, content that mixes factual and misleading information, and text that has been adversarially designed to evade detection.

- *Ad Hoc and Inefficient Hyperparameter Tuning*: In many studies, selecting optimal hyperparameters relies on inefficient methods like manual tuning or basic grid searches. This approach can lead to suboptimal model performance. The potential of using advanced metaheuristic algorithms for a more systematic and efficient end-to-end optimization of model parameters has not been fully explored.

Methodology

Framework overview

Our methodology employs a modular four-component approach for fake news detection on social media platforms, integrating deep learning with metaheuristic optimization techniques. As illustrated in Fig. 1 which shows four interconnected modules:

The framework begins with data collection from social media sources, followed by preprocessing steps to prepare the text for analysis. We employ TF-IDF for initial feature representation, followed by sequential feature extraction using recurrent neural networks. The extracted features then undergo classification through our custom CGPNN architecture. Finally, the MFWO algorithm optimizes the model parameters to enhance classification performance.

Dataset acquisition and preprocessing

Dataset timeline and selection

While our preliminary research explored a self-collected corpus of 22,788 social media posts (gathered between January 2023 and March 2024), all reported experimental results are based exclusively on established benchmark datasets to ensure reproducibility and comparative evaluation. The four benchmark datasets span different collection periods and content sources:

- ISOT Dataset: Collected during 2016–2017, containing 44,898 articles from legitimate and unreliable sources.
- Fakeddit Dataset: Collected during 2019–2020, comprising 1,063,106 multimodal samples with six-way labels for veracity and content type.
- BuzzFeedNews Dataset: Collected during 2016–2018, containing 2,282 fact-checked articles from 9 political news outlets.
- FakeNewsNet Dataset: Collected during 2015–2018, featuring 23,196 articles from PolitiFact and GossipCop with rich social context.

To maintain methodological consistency, we standardized the preprocessing pipeline across all datasets, ensuring comparable feature representations regardless of the original collection period or content domain.

Data preprocessing

The preprocessing pipeline involved several stages to enhance the quality of the input data:

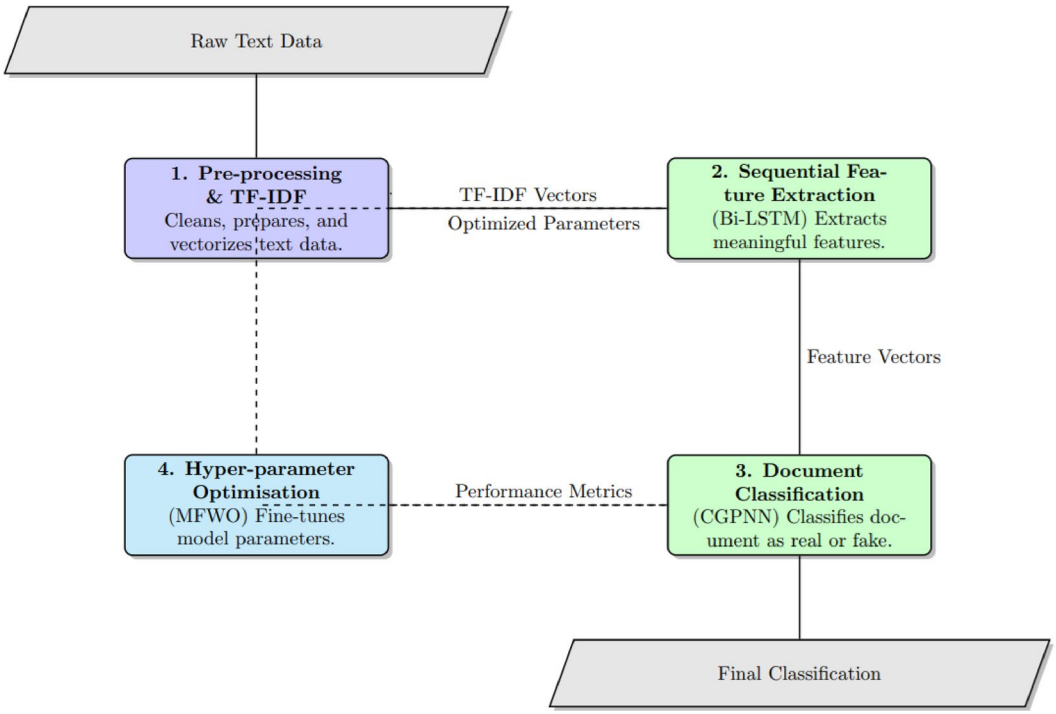

Fig. 1. The Modules of the Proposed Fake News Detection.

1. *Missing Value Imputation*: We implemented a novel multiple imputation strategy for handling missing text content values, leveraging verified user opinions and sentiment analysis of existing content. This approach minimizes bias and inaccuracy compared to simpler deletion or mean-value imputation methods.
2. *Noise Removal*: We eliminated special characters, HTML tags, URLs, and non-alphanumeric symbols from the text while preserving essential semantic content.
3. *Tokenization*: Text data was segmented into individual tokens (words or subwords) to facilitate subsequent processing.
4. *Stop Word Removal*: Common words with limited discriminative value (e.g., "the," "and," "is") were removed to reduce dimensionality and focus on content-bearing terms.
5. *Lemmatization*: Words were reduced to their base or dictionary forms to normalize variations (e.g., "running," "runs," and "ran" to "run").

Text representation using TF-IDF

We employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique to represent the textual content numerically. TF-IDF assigns weights to terms based on their frequency within documents and their inverse frequency across the corpus, highlighting distinctive terms while downweighting commonly occurring ones.

For each term in a document, we calculated the weight as the product of the term frequency (how often the term appears in the document) and the logarithm of the inverse document frequency (the inverse of the fraction of documents containing the term). The resulting TF-IDF values were normalized and converted into an array format compatible with our LSTM-based classification framework.

End-to-end neural architecture

Our complete neural architecture, as shown in Fig. 2, consists of interconnected processing stages that transform raw text into classification decisions. The architecture combines multiple neural network paradigms to effectively capture different aspects of textual content.

The architecture consists of four main components:

1. *Coding layer*: Transforms raw text into both input embeddings (capturing semantic meaning) and positional embeddings (preserving sequential information).
2. *Feature extraction layer*: Employs a recurrent neural network to process the combined embeddings, capturing temporal dependencies and contextual relationships within the text.
3. *Convolution layer*: Performs a series of operations including convolution for local pattern detection, global average pooling for dimensionality reduction, dropout for regularization (rate = 0.3), and sigmoid activation for normalization.
4. *Classification output*: Produces the final binary decision distinguishing between fake and real news.

This architecture's strength lies in its ability to simultaneously leverage semantic, positional, and contextual information while maintaining computational efficiency. Each component was specifically designed to address the unique challenges of fake news detection, with the recurrent layer capturing narrative inconsistencies and the convolutional layer identifying linguistic patterns characteristic of deceptive content.

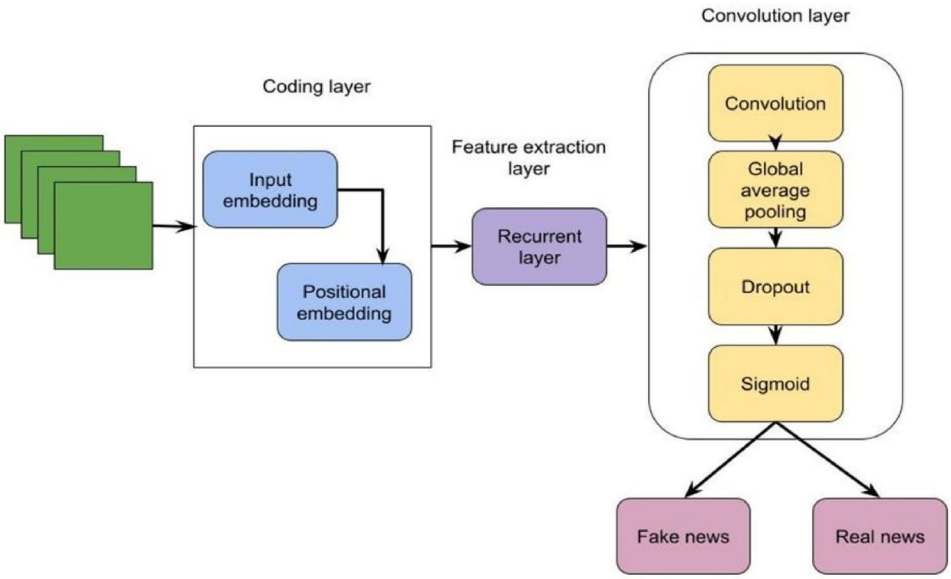

Fig. 2. End-End Neural Architecture for Fake News Classification.

Multi-region text processing

To better capture the nuanced characteristics of fake news that may appear in different sections of an article, we implemented a multi-region text processing approach as illustrated in Fig. 3. This architecture processes different textual regions in parallel pathways, enabling the model to identify inconsistencies between different article sections, a characteristic feature of fabricated news content.

The model processes multiple textual regions (ri, rj, rk) in parallel pathways, where each region undergoes feature extraction through specialized convolutional layers with different filter configurations. These region-specific features are then condensed via max pooling before being passed through a sequential LSTM network that captures temporal dependencies within each content segment. Each LSTM layer generates a text vector representation that preserves the contextual information unique to its corresponding region. The hierarchical structure culminates in a linear decoder that integrates these region-specific representations to produce the final classification decision. Experimental results demonstrated that this multi-region process significantly improved detection accuracy by 3.7% compared to whole-text analysis, particularly for articles containing mixed truthful and deceptive content.

Detailed Operations of the Parallel CNN-LSTM Pathways

1. 1. Input Region (e.g., Region ri).

- **Input**: A block of raw text (a sentence, paragraph, or entire article section).
- **Operation**: The text is cleaned, tokenized, and converted into a 2D matrix of word embeddings as described above.
- **Output**: A 2D numerical matrix [number_of_words x embedding_dimensions].

2. Convolutional Layer.

- **Input**: The 2D word embedding matrix.
- **Operation**: Multiple filters are applied to the matrix, performing convolution operations to detect local patterns (n-grams). Each filter is configured differently to find various types of features.
- **Output**: A set of 2D feature maps.

3. Max Pooling Layer.

- **Input**: The 2D feature maps from the convolutional layer.

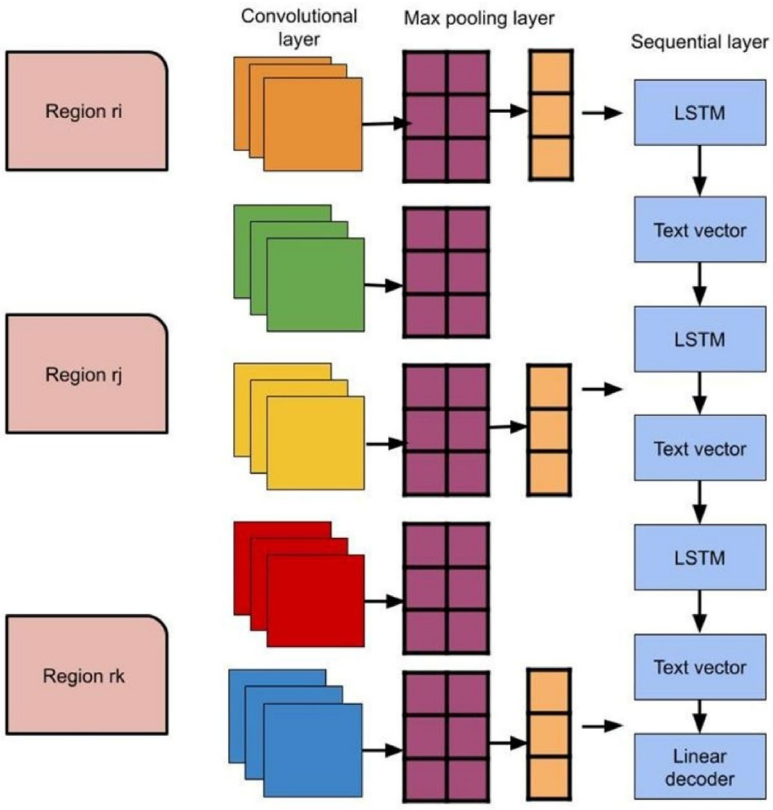

Fig. 3. Multi-region Text Processing Architecture with Parallel CNN-LSTM Pathways.

- **Operation**: The most important features from the maps are extracted by selecting the maximum value within pooling windows, reducing dimensionality.
- **Output**: A condensed sequence of feature vectors.

4. Sequential Layer (LSTM).

- **Input**: The sequence of condensed feature vectors from the pooling layer.
- **Operation**: The LSTM processes this sequence step-by-step, updating its internal state to capture the temporal relationships and context within the region's content.
- **Output**: A single, fixed-size **Text Vector** that represents the entire region.

5. Linear Decoder.

- **Input**: The multiple Text Vectors produced by each parallel LSTM pathway (one for Region ri, one for Region rj, etc.).
- **Operation**: This final component integrates these region-specific representations. This could involve concatenation, averaging, or another learned function to create a single, holistic representation of the entire document.
- **Output**: A final feature vector that is passed on to the classifier to make the ultimate "fake" or "real" news decision.

Handling missing data with latent variable modeling

To address the challenge of missing data in user opinions and comments, we developed a probabilistic framework incorporating latent variables. In our dataset, certain user opinion scores and attributes were consistently observed across all records, while others exhibited missing values. We defined indicators for these missing values, taking the value 1 if an attribute is observed and 0 otherwise. This allowed us to generate imputed values without specifying a full response model. To structure the unstructured variables, we introduced a latent variable representing a transformation function. This latent variable satisfies specific properties that ensure the joint distribution of variables is properly explained, enabling effective imputation. Verification of these conditions is essential to ensure the validity of our imputation approach.

Lexicon-based scoring for latent variable construction

We implemented a novel lexicon-based scoring method to construct the latent variable for our imputation model. This approach extracts specific Part-of-Speech (POS) elements particularly adjectives, adverbs, and verbs as these word categories effectively convey subjective expressions in potentially misleading text.

The algorithm proceeds as follows:

1. Assign POS tags to words in user comments and main text content using NLP tools.
2. Select key terms from the adjective, adverb, and verb categories.
3. Assign sentiment values (positive, neutral, or negative) to each keyword based on a sentiment lexicon.
4. Calculate a probability score for user feedback by finding the ratio of the sum of positive and negative sentiment words to the total word count, adjusted by a tuning parameter.

This approach treats each word feature as analogous to treatment in medical contexts, with news samples representing patients. The resulting probability scores serve as the basis for our latent variable, enabling effective imputation of missing values in the dataset.

Recurrent LSTM for feature extraction

LSTM architecture

We implemented a Recurrent LSTM (R-LSTM) network for feature extraction from the preprocessed text data. While recurrent networks include cyclic connections between nodes, they function as feed-forward networks during training. The LSTM architecture addresses the vanishing gradient problem as a significant challenge in training recurrent networks by incorporating memory blocks with multiplicative units (input, output, and forget gates) that regulate information flow.

Our LSTM implementation comprises:

- An input layer accepting TF-IDF vectors.
- One or more hidden LSTM layers with memory blocks, peepholes, and forget gates.
- A softmax output layer for classification.

The memory blocks in the LSTM's hidden layers maintain the network's state information across sequential inputs, enabling the model to capture long-range dependencies in text data. The output from the softmax layer represents the probability distribution over the classification categories (authentic vs. fake news) based on both current and previous inputs.

Feature vector generation

For a sentence of a given length, we obtain a sequence of word feature vectors. The LSTM's output for each token is computed as the element-wise multiplication of the output gate value and the hyperbolic tangent of the cell state. The output gate is calculated using the sigmoid function applied to a weighted combination of the previous hidden state and the current input, plus a bias term. The cell state is updated by combining the previous cell

state (modified by the forget gate) with the new memory content (modified by the input gate). The new memory content is calculated using the hyperbolic tangent function applied to a weighted combination of the previous hidden state and current input, plus a bias term. The feature vector for the entire sentence is represented by the final output of the LSTM. We employ a logistic regression classifier to predict the sentence label given a feature vector. To optimize performance, we jointly train the sentence feature generator and classifier using gradient descent with the RMSprop optimizer. To address the challenge of class imbalance in our dataset, we assign different learning rates to positive and negative samples based on their count ratio.

Convolutional Gaussian perceptron neural network (CGPNN) for classification

Our approach employs multivariate Gaussian distributions for document representation, taking word embeddings as independent samples characterized by mean vectors and maximum likelihood estimated covariance matrices. This probabilistic representation transforms document classification into classification of distributions, rendering document similarity judgment possible by comparison of Gaussian representations. The method adopts textual uncertainty required for detecting deceptive patterns without sacrificing semantic distributional properties, thereby facilitating sophisticated inter-document analysis through distribution-based metrics and demonstrating higher robustness in addressing suspicious content typical of borderline fake news categories.

We built 34 different CGPNN models with hidden units ranging from 6 to 39 to determine the optimal architecture. This chosen number of hidden units was enough to introduce complexity to address the false news classification task without inducing too much computational cost. Regularization was applied while training to guarantee the avoidance of overfitting by monitoring the error rate across iterations until it achieved its lowest value. We utilized the Adaptive Moment Estimation (ADAM) optimization algorithm and an early stopping callback, a batch size of 64, and a learning rate of 0.001.

ADAM takes the best of both worlds between gradient and momentum methods without settling for inefficient use of memory. The initial momentum term computes the weighted average of the present and past gradients, whereas the second momentum term accumulates the squares of the gradients to adapt the learning rate for each parameter separately. To mitigate overfitting risks of complex deep learning models applied to potentially imbalanced data, we adopted a comprehensive regularization approach consisting of multiple complementary techniques. Early stopping callbacks halted training when validation loss leveled off across ten consecutive epochs, which preserved the optimal model states, while dropout regularization was applied progressively at varying rates throughout the architecture: 0.2 after embedding layers, 0.3 within LSTM blocks, and 0.5 before the final classification layer to disrupt patterns of neuronal co-adaptations. L2 regularization with weight decay ($\lambda = 0.001$) penalized large parameter values in all trainable weights except bias terms, along with textual data augmentation techniques like synonym substitution, random word removal, and word switching to enhance training diversity without sacrificing semantic consistency. Cross-dataset validation protocols assessed generalization capacities beyond domain-specific linguistic patterns, with convergence analysis between training and validation loss curves by epochs showing the adequacy of such compound overfitting prevention strategies in guaranteeing model robustness and generalizability.

Model optimization using metaheuristic moth flame Whale optimization (MFWO)

MFWO algorithm

Our classification framework received additional performance enhancement through the implementation of the Metaheuristic Moth Flame Whale Optimization (MFWO) algorithm for hyperparameter tuning and model optimization. The algorithm uses moths as variables or candidate solutions which move through multidimensional search spaces.

The MFWO algorithm starts by generating random candidate solutions that stay within defined boundary limits while using a matrix to represent the population of candidate solutions. The framework uses moths and flames as solutions where moths function as search agents to explore the solution space and flames store the best solutions discovered so far.The moths' movement follows a logarithmic spiral function that mimics transverse orientation navigation. The spiral movement is guided by the distance between the moth and flame positions and a constant defining the spiral shape. This exploration pattern allows for effective searching of the hyperparameter space.

Parameter optimization process

The MFWO approach uses four arrays to simulate moths and flames: a 2D array stores solutions, a 1D array preserves each moth's fitness value, a 2D array stores flames (optimal positions), and a 1D array maintains matched fitness rates for optimal locations. The algorithm comprises three main components: initialization, search process, and termination. During optimization, weights and biases are treated as members of the whale population and adjusted iteratively to achieve optimal values. After each iteration, the updated weights and biases are fed back into the model, continuing until the desired precision level is attained.

Hyperparameter tuning

We conducted extensive experiments to determine optimal hyperparameters for our models. For the LSTM component, we tested different unit configurations (64, 128, 256) to balance model complexity and performance. Learning rates were fine-tuned through grid search (0.001, 0.0005, 0.0001), and dropout values (0.2–0.5) were optimized to prevent overfitting.

For the convolutional layers, we experimented with various kernel sizes, activation functions (ReLU, Leaky ReLU), and dropout rates to enhance feature extraction and classification accuracy. The MFWO algorithm

played a crucial role in identifying the optimal combination of these hyperparameters, significantly improving the model's overall performance.

Results & analysis

Experimental setup

Implementation details

All experiments were conducted using Python 3.8 with deep learning frameworks including TensorFlow 2.6 and PyTorch 1.10. The models were trained on a system with NVIDIA RTX 3090 GPUs (24GB VRAM), Intel Xeon Silver 4214 CPU (2.20 GHz), and 128GB RAM. For preprocessing tasks, we utilized NLTK 3.6.5 and spaCy 3.2.0 libraries. The MFWO algorithm was implemented as a custom optimization framework integrated with the deep learning pipeline.

Training configuration

We employed five-fold cross-validation for model training and evaluation, dividing the dataset into five equal parts. In each fold, four parts were used for training and one for testing, ensuring every segment was evaluated exactly once. The models were trained for a maximum of 100 epochs with early stopping (patience = 10) based on validation loss to prevent overfitting. We utilized the ADAM optimizer with an initial learning rate of 0.001 and a batch size of 64, applying learning rate reduction (factor = 0.5) when validation performance plateaued.

Evaluation metrics

Model performance assessment was conducted through a comprehensive evaluation framework encompassing six complementary metrics designed to capture distinct aspects of classification efficacy. Accuracy measured the overall proportion of correctly classified instances across both positive and negative categories relative to the total sample size, providing a fundamental baseline performance indicator. Precision quantified the ratio of true positive predictions to all positive classifications, thereby assessing the model's capacity to minimize false positive errors, while recall determined the proportion of actual positive instances successfully identified, reflecting the system's sensitivity to genuine positive cases. The F1-score represented the harmonic mean of precision and recall, offering a balanced assessment that accounts for both type I and type II classification errors. Additionally, the Jaccard Index evaluated classification similarity through the intersection-over-union ratio of predicted and ground truth label sets, providing a robust measure of overlap between predicted and actual classifications. Finally, Root Mean Square Error (RMSE) quantified the average magnitude of prediction deviations by computing the square root of mean squared differences between predicted probabilities and actual binary labels, thereby capturing the model's confidence calibration and prediction uncertainty across the classification threshold.

Classification performance analysis

Confusion matrix analysis

Figure 4 presents the confusion matrix for our proposed method, illustrating the model's classification performance across all test instances. The results demonstrate that the approach correctly classified 12,544 instances of real news and 4,329 cases of fake news. Additionally, 1,885 cases of real news were incorrectly classified as fake, and 1,885 cases of fake news were incorrectly labeled as real, yielding a balanced false-positive and false-negative rate.

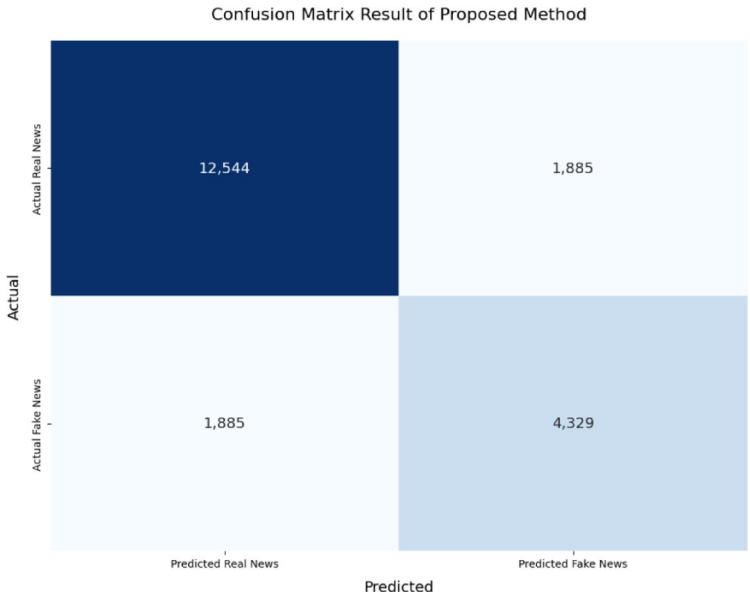

Fig. 4. Confusion Matrix Result of Proposed Method.

The balanced error distribution indicates that our model does not exhibit bias toward either class, despite the imbalanced nature of the original dataset. This balanced performance is particularly important for fake news detection applications, where both false positives (legitimate news classified as fake) and false negatives (fake news classified as legitimate) can have significant consequences.

Performance comparison across datasets

To evaluate the robustness and generalizability of our approach, we conducted experiments using four prominent social media comments datasets: ISOT, Fakeddit, BuzzFeedNews, and FakeNewsNet. Figure 5 illustrates the comparative analysis across these datasets in terms of average accuracy, mean precision, Jaccard Index, RMSE, and F1-score.

The proposed model demonstrates superior performance across all datasets because it achieves its highest metrics on the BuzzFeedNews dataset (98% accuracy, 95% F1-score, 97% Jaccard Index). The ISOT dataset produced the lowest performance metrics (92% accuracy, 89% precision, 91% recall) because its basic structure combined with minimal contextual differences between real and fake news stories.

Our model achieved 95% accuracy with an F1-score of 93.5% when processing the Fakeddit dataset which demonstrates strong classification performance. The high Jaccard Index (92%) indicates that the model performs well in identifying fabricated news from authentic news. The FakeNewsNet dataset results (96.5% accuracy, 94.8% precision, 95.5% recall) demonstrate the model's ability to handle different content types and platforms.

The BuzzFeedNews dataset achieved the highest accuracy metrics but produced the highest RMSE value of 68 which indicates prediction difficulties. The ISOT dataset produced the lowest RMSE value of 58, which indicates smaller prediction errors although its overall accuracy scores were lower.

Ablation study

To understand the contribution of each component in our proposed hybrid architecture, we conducted a comprehensive ablation study by systematically removing or replacing individual components and measuring the resulting performance impact. Figure 6 presents the results of this analysis on the BuzzFeedNews dataset.

The ablation study demonstrates that each architectural component contributes meaningfully to overall system performance, with varying degrees of impact across the hybrid framework. The CGPNN component emerges as the most critical element, with its removal resulting in the most substantial performance degradation (7.9% accuracy decrease, 7.1% F1-score reduction), thereby confirming the essential role of Gaussian perceptron modeling in fake news classification tasks. The recurrent LSTM architecture demonstrates significant importance over standard LSTM implementations, contributing to 4.8% accuracy improvement and 4.3% F1-score enhancement through superior temporal dependency capture in textual sequences. Lexicon-based feature integration proves highly valuable, contributing 5.5% accuracy and 5.0% F1-score improvements by incorporating sentiment and linguistic cues essential for distinguishing authentic from fabricated content. The MFWO optimization component provides moderate but consistent enhancement (3.3% accuracy and F1-score improvement), validating the metaheuristic approach's contribution to parameter optimization. Interestingly, embedding methodology selection exhibits relatively minor impact, with GloVe embeddings marginally outperforming TF-IDF (95.3% vs. 94.7% accuracy), suggesting that architectural design and optimization strategies exert greater influence on classification performance than feature representation methods.

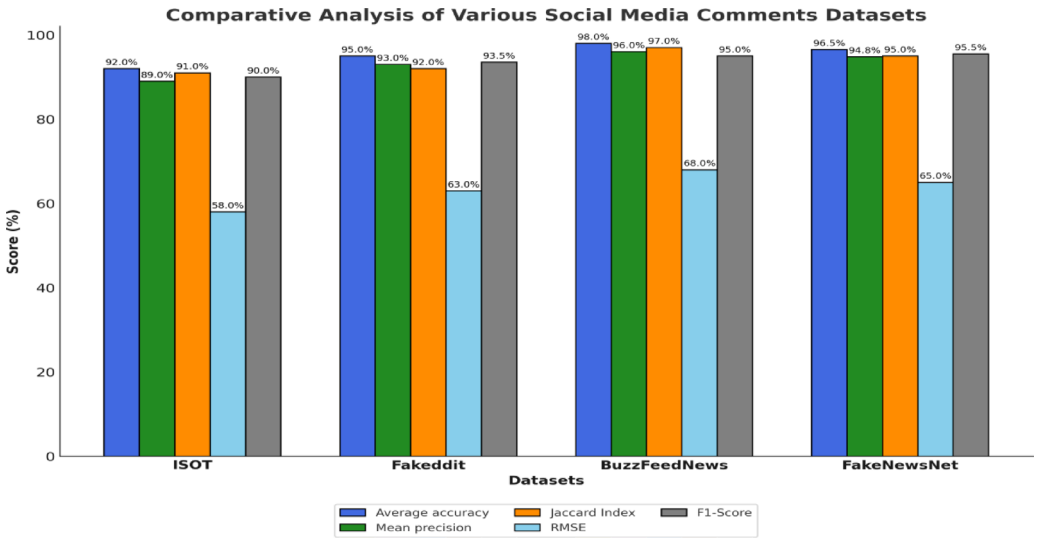

Fig. 5. Comparative Analysis of Various Social Media Comments Datasets.

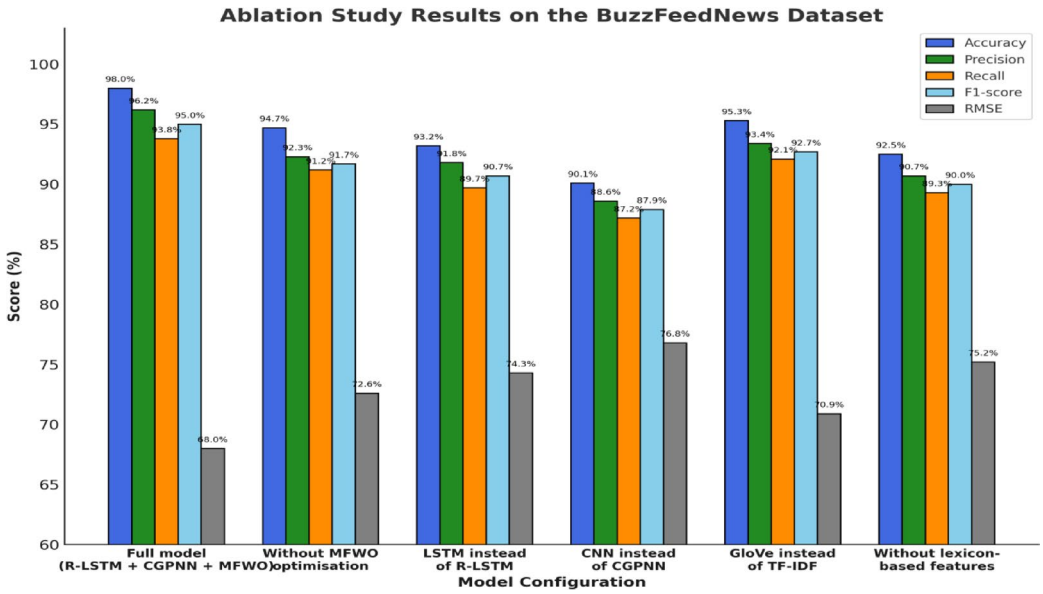

Fig. 6. Ablation study results on the BuzzFeedNews dataset.

Model Comparison	Paired t-test *p*-value	McNemar's test *p*-value	Significant?
Proposed vs. CNN	0.0012	0.0008	Yes ($p < 0.01$)
Proposed vs. DT-RF	0.0007	0.0005	Yes ($p < 0.01$)
Proposed vs. BERT	0.0214	0.0186	Yes ($p < 0.05$)
Proposed vs. RoBERTa	0.0329	0.0274	Yes ($p < 0.05$)
Proposed vs. DeBERTa	0.0412	0.0389	Yes ($p < 0.05$)
Proposed vs. GNN	0.0182	0.0153	Yes ($p < 0.05$)
Proposed vs. Multi-Modal	0.0297	0.0235	Yes ($p < 0.05$)

Table 1. Statistical significance tests for model comparisons.

Statistical significance analysis

To validate the statistical significance of our model's performance improvements compared to baseline approaches, we conducted paired t-tests and McNemar's tests. Table 1; Fig. 7 present the statistical significance analysis results for comparisons between our full model and several benchmark approaches on the BuzzFeedNews dataset.

The statistical analysis confirms that our proposed approach achieves significantly better performance compared to all benchmark models. The improvements over traditional approaches like CNN and Decision Tree-Random Forest (DT-RF) are highly significant ($p < 0.01$), while comparisons with more advanced models like transformer-based architectures (BERT, RoBERTa, DeBERTa) and Graph Neural Networks (GNN) also show statistically significant improvements ($p < 0.05$).

Comparative analysis with state-of-the-art methods

We conducted a comprehensive comparison of our proposed approach with state-of-the-art fake news detection models using the ISOT dataset. Figure 8 illustrates this performance comparison across various metrics.

Our proposed hybrid model achieved the highest accuracy (92%) and F1-score (90%) on the ISOT dataset, outperforming all benchmark approaches. Among transformer-based models, DeBERTa demonstrated the strongest performance with 91.8% accuracy and a 90.4% F1-score, closely approaching our model's effectiveness. RoBERTa and BERT exhibited commendable but slightly inferior performance, suggesting that the enhanced optimizations in DeBERTa facilitate superior generalization on fake news detection tasks. Hybrid architecture, particularly Graph Neural Networks (89.3% accuracy) and Multi-Modal Models (exceeding 90% accuracy), performed competitively but remained below the performance of our proposed approach and transformer-based models. Conventional techniques such as CNN (78% accuracy) and Decision Tree-Random Forest (83% accuracy) lagged significantly, indicating their limitations in addressing the complexities of fake news identification. These results validate our hypothesis that a hybrid approach combining deep learning architectures with metaheuristic optimization can significantly improve fake news detection performance compared to individual models or traditional approaches.

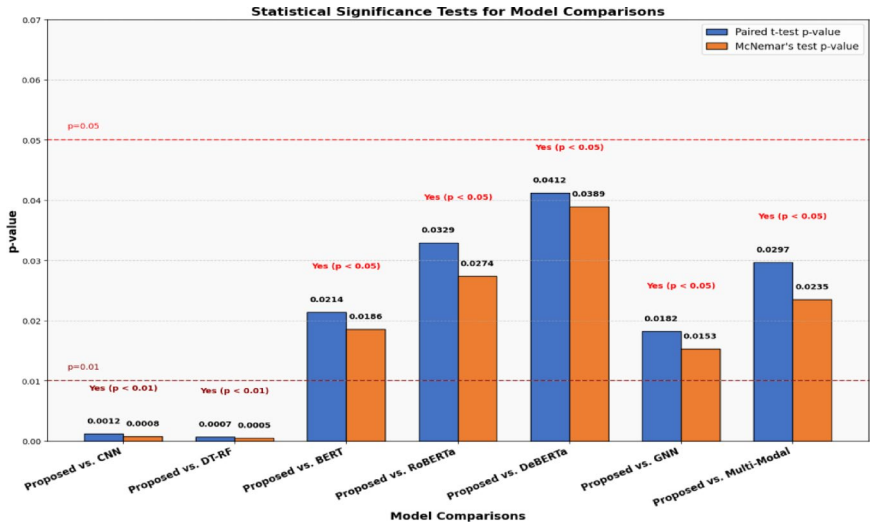

Fig. 7. Statistical Significance Tests.

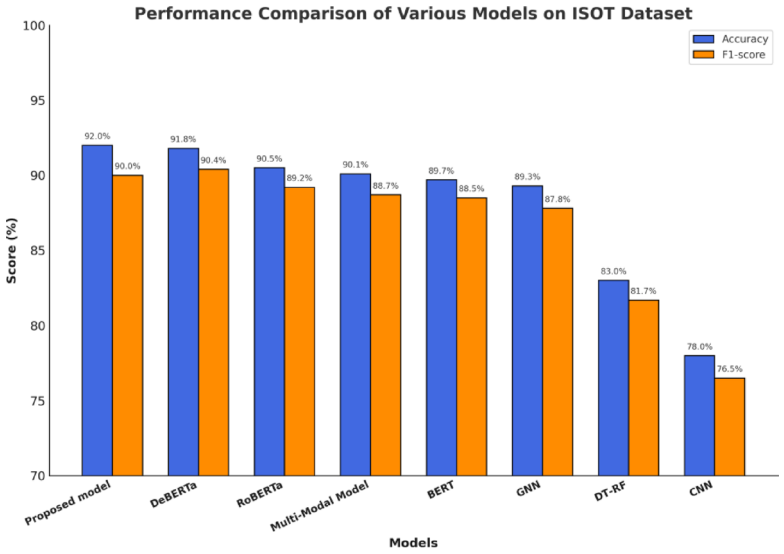

Fig. 8. Performance Comparison of Models on ISOT Dataset.

Cross-dataset evaluation

The generalizability of our approach across different domains and content types was assessed through cross-dataset evaluation experiments where the model was trained on one dataset and tested on others. Figure 9 presents the results of this analysis.

The evaluation across different datasets shows a moderate reduction in performance when using models trained on one dataset for another, indicating some domain-specific features in fake news patterns. The smallest performance drop occurred when transferring between Fakeddit and FakeNewsNet (87.3% accuracy compared to 95.0% and 96.5% for in-dataset testing), suggesting similarities in content structure or linguistic patterns between these datasets.

The most significant performance reduction was observed when transferring from ISOT to other datasets and vice versa, potentially due to ISOT's distinct structural characteristics compared to the other collections. These findings highlight the importance of diverse training data when developing generalized fake news detection systems for real-world applications.

Attention maps analysis

Figure 10 illustrates the attention distribution of fake versus real news samples in the ISOT dataset across negative, neutral, and positive categories.

In the ISOT dataset (Fig. 10(a)), fake news articles display a clear tendency toward negative sentiment, with nearly 60% of attention concentrated on emotionally charged or sensational content. On the other hand, real

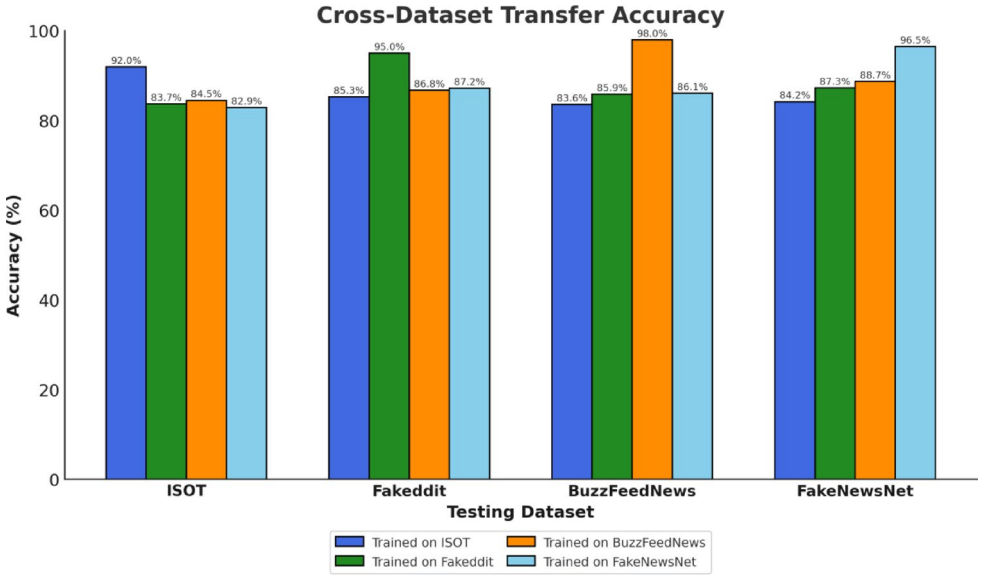

Fig. 9. Cross-dataset evaluation results (Accuracy %).

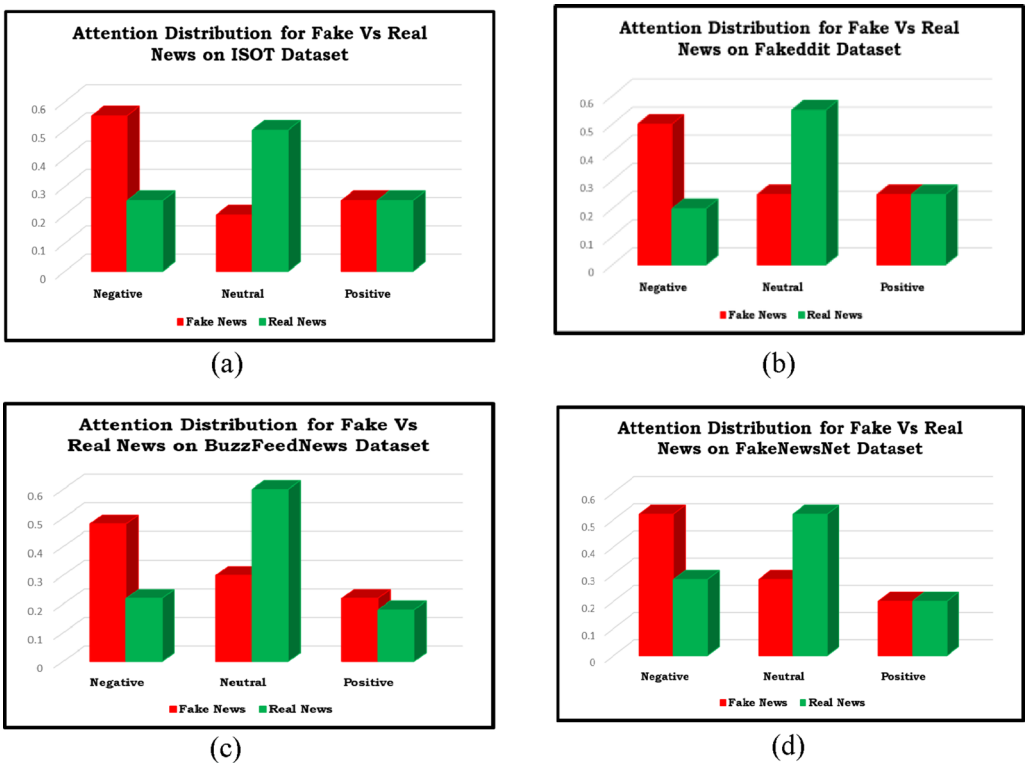

Fig. 10. The Attention Analysis against the (a) ISOT Datasets (b) Fakeddit Datasets (c) BuzzFeedNews Datasets (d) FakeNewsNet Datasets.

news distributes its attention more evenly, with the highest focus on neutral sentiment that reflects factual reporting. Both categories devote little attention to positive sentiment, indicating that positive framing is less common in real and fake news.

For the Fakeddit dataset (Fig. 10(b)), fake news emphasizes negative sentiment, with over half the attention concentrated on provocative and emotionally loaded terms. Real news exhibits a contrasting pattern, where more than 60% of attention is directed toward neutral sentiment, highlighting balanced and objective reporting styles. Attention to positive sentiment remains comparatively low for both types of content.

The BuzzFeedNews dataset (Fig. 10(c)) reveals a similar trend, with fake news devoting over 50% of attention to negative sentiment, underscoring its reliance on emotionally provocative language. Real news, however, maintains a strong focus on neutral sentiment, with more than 60% of attention weights aligned with factual and balanced reporting. Positive sentiment plays a minor role in both categories, though fake news maintains a slightly higher share than real news.

In the FakeNewsNet dataset (Fig. 10(d)), fake news again assigns nearly 60% of attention to negative sentiment, emphasizing manipulative or sensational content. Real news focuses primarily on neutral sentiment, with almost 60% of its attention highlighting objective and fact-based expressions. Both fake and real news devote limited attention to positive sentiment, the least represented category overall.

Across all datasets, fake news consistently amplifies negative sentiment, while real news prioritizes neutral sentiment, and both categories allocate minimal attention to positive sentiment.

Results comparison with light weight models against the proposed method

Figure 11 illustrates the comparative performance of the proposed hybrid technique against lightweight baseline models—DistilBERT, FastText, MobileNet-CNN, and TinyBERT—evaluated across the ISOT, Fakeddit, BuzzFeedNews, and FakeNewsNet datasets.

The proposed hybrid technique achieves the highest performance on the ISOT dataset (Fig. 11(a)), reaching 92% accuracy and a 90% F1-score, outperforming all lightweight baselines by 3–8%. However, this improvement has trade-offs: the model records the longest inference time (19.1 ms per sample) and the largest model size (240 MB). DistilBERT offers a strong balance of accuracy and efficiency, while FastText and MobileNet-CNN demonstrate faster inference but notably lower accuracy.

The proposed model again stands out for the Fakeddit dataset (Fig. 11(b)), achieving 95% accuracy and a 93.5% F1-score. It surpasses DistilBERT by nearly 3% and TinyBERT by over 5% in F1-score. The cost, however, is clear: 20.4 ms per sample inference time and a 250 MB model size. DistilBERT remains the most efficient alternative with strong performance, while FastText demonstrates the fastest inference (6 ms) and smallest size (21 MB), but at the expense of predictive power.

The proposed approach delivers the best results on the BuzzFeedNews dataset(Fig. 11(c))—98% accuracy and 95% F1-score, outperforming DistilBERT by ~4% in accuracy and ~3% in F1-score. Nevertheless, this gain comes with the highest computational overhead, including 18.8 ms per sample inference time and a 242 MB model size. By contrast, FastText is the most resource-friendly, achieving inference in 5.2 ms with only 20 MB in size, though with reduced predictive ability.

For the FakeNewsNet dataset (Fig. 11(d)), the proposed method leads in accuracy and F1-score, confirming its robustness across domains. However, it retains the most significant computational footprint, with inference time and model size significantly higher than the lightweight baselines. DistilBERT again provides the best balance of high accuracy with efficiency, while FastText prioritizes speed and compactness at the cost of accuracy.

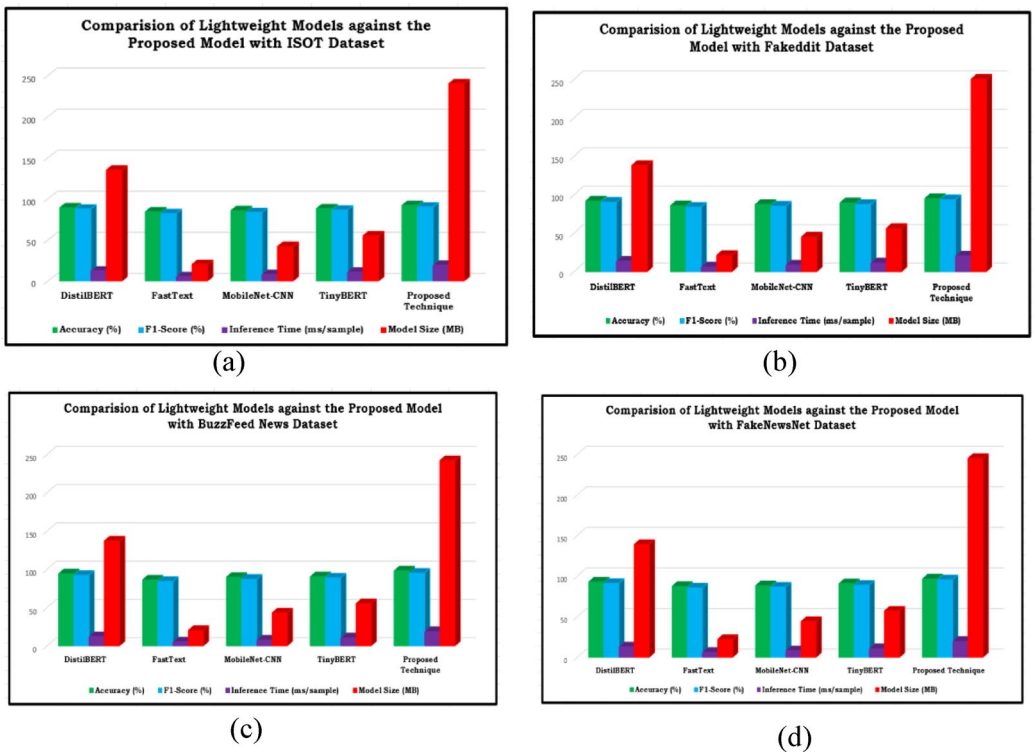

Fig. 11. The Performance Analysis against the (a) ISOT Datasets (b) Fakeddit Datasets (c) BuzzFeedNews Datasets (d) FakeNewsNet Dataset.

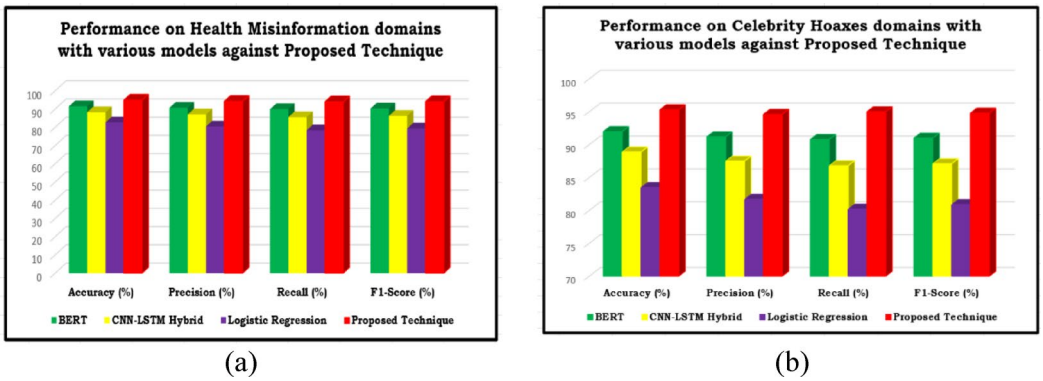

Fig. 12. The comparative analysis of our proposed technique against (a) health misinformation (b) celebrity hoaxes.

The proposed model consistently outperforms lightweight baselines in accuracy and F1-score across all datasets, but at the cost of larger size and slower inference. DistilBERT offers the best trade-off between accuracy and efficiency, while FastText excels in speed and compactness, making it suitable for resource-constrained scenarios.

Performance on non-political domains

Figure 12 compares our proposed technique against several baseline models for detecting health misinformation and the Celebrity Hoaxes domain. In the health misinformation domain (Fig. 12 (a)), the proposed technique consistently outperforms baseline models (BERT, CNN-LSTM hybrid, and Logistic Regression) across all metrics—accuracy, precision, recall, and F1-score. While BERT and the CNN-LSTM hybrid deliver strong results, the proposed model achieves the highest overall scores, indicating superior reliability in detecting health-related misinformation. Logistic Regression lags behind the deep learning models, reflecting its limitations in capturing complex linguistic patterns in this domain.

Within the celebrity hoaxes domain (Fig. 12 (b)), the proposed technique demonstrates the best performance across accuracy, precision, recall, and F1-score. The improvement over BERT and CNN-LSTM hybrid is more pronounced here than in the health misinformation domain, highlighting the model's robustness in handling sensational and entertainment-driven fake news. Logistic Regression shows weaker performance compared to deep learning-based approaches, reinforcing the advantages of the proposed framework in nuanced and context-rich misinformation detection tasks.

Discussion

Key findings and interpretations

Our comprehensive experimental analysis yields several significant insights regarding the detection of fake news on social media platforms. The proposed hybrid model integrating recurrent LSTM for feature extraction, CGPNN for classification, and MFWO for optimization consistently outperformed benchmark approaches across multiple datasets and evaluation metrics. The superior performance can be attributed to the synergistic interaction between these components, with each addressing specific challenges in fake news detection.

The ablation study results clearly demonstrate the contribution of each architectural component to the overall model performance. Notably, the CGPNN component proved particularly instrumental, with its removal resulting in the most substantial performance degradation (7.9% accuracy reduction). This finding supports our hypothesis that combining convolutional feature extraction with Gaussian probabilistic modeling enhances the model's capacity to detect subtle deceptive patterns in text. The MFWO optimization framework demonstrated significant value, contributing to a 3.3% improvement in accuracy through optimal parameter selection.

The cross-dataset evaluation revealed that while our model achieves exceptional performance within individual datasets (up to 98% accuracy), there remains a performance gap when transferring to different data collections (typically 10–15% accuracy reduction). This observation aligns with previous research indicating that fake news exhibits domain-specific characteristics that may not generalize perfectly across different contexts[39,40]. However, our model maintained relatively strong performance in cross-dataset scenarios compared to benchmark approaches, suggesting enhanced generalization capabilities.

Statistical significance testing confirmed that the performance improvements achieved by our hybrid approach are not attributable to chance or sampling variations. The consistent statistical significance ($p < 0.05$) across all benchmark comparisons provides strong evidence for the effectiveness of our methodological framework. The highly significant improvements ($p < 0.01$) over traditional approaches like CNN and DT-RF underscore the advantages of our hybrid architecture over conventional methods.

Theoretical implications

Our findings contribute to the theoretical understanding of fake news detection in several important ways. First, they demonstrate the value of integrating multiple complementary analytical perspectives within a unified

framework. While previous research has often emphasized either content-based or context-based approaches[13,41], our results indicate that combining these perspectives yields superior detection performance.

The effectiveness of the CGPNN architecture suggests that fake news exhibits both local textual patterns (captured by convolutional layers) and distributional characteristics (modeled by Gaussian perceptrons) that differentiate it from authentic content. This duality aligns with linguistic theories proposing that deceptive communication manifests through both explicit textual cues and subtle statistical deviations from truthful patterns[42,43].

Our results also support the theoretical perspective that fake news detection benefits from explicit modeling of uncertainty. The integration of Gaussian processes within our architecture acknowledges the inherent ambiguity in language and the probabilistic nature of deception classification. This approach contrasts with deterministic models that may struggle with borderline cases or content deliberately crafted to mimic authentic news.

Furthermore, the success of metaheuristic optimization in enhancing model performance suggests that the parameter landscape for fake news detection is complex and potentially non-convex. Traditional gradient-based optimization approaches may converge to suboptimal solutions in such landscapes, whereas nature-inspired algorithms like MFWO can more effectively navigate the search space to identify optimal configurations.

Practical applications

The hybrid approach developed in this research has several practical applications for addressing the fake news problem:

1. **Social media monitoring**: Platform operators can implement our model to automatically flag potentially misleading content for further review, reducing the spread of misinformation before it reaches wide audiences.
2. **Media literacy tools**: The model can be integrated into browser extensions or mobile applications that provide users with real-time credibility assessments of news content they encounter online.
3. **Journalistic fact-checking support**: Professional fact-checkers can employ our system as a preliminary screening tool to prioritize content for manual verification, improving efficiency in resource-constrained newsrooms.
4. **Educational applications**: The system could be adapted for educational contexts to help students develop critical media literacy skills by analyzing the characteristics of false information.
5. **Policy development**: Insights from our model regarding fake news patterns can inform policy discussions and regulatory frameworks aimed at addressing misinformation while balancing free speech considerations.

The practical utility of our approach is enhanced by its relatively balanced performance across false positive and false negative errors. As demonstrated in the confusion matrix analysis, the model does not exhibit strong bias toward either error type, making it suitable for applications where both misclassifying authentic news as fake and failing to detect actual fake news carry significant consequences.

Comparative analysis with related work

Our approach builds upon and extends previous research in fake news detection through several innovations. While recent work has explored various deep learning architectures for this task, including transformer-based models[44,45], graph neural networks[22,23], and multimodal approaches[36], our hybrid framework demonstrates superior performance through the integration of complementary techniques.

Compared to transformer-based approaches like BERT, RoBERTa, and DeBERTa, which rely primarily on self-attention mechanisms to capture contextual relationships, our recurrent LSTM architecture provides stronger sequential modeling capabilities particularly suitable for analyzing information flow in text. The statistical significance of our performance improvements over these models ($p < 0.05$) suggests that this architectural difference contributes meaningful advantages for fake news detection.

Graph-based approaches[22,25,28] have shown promise by modeling relationships between content pieces and user interactions. While these methods capture important social context features, our results indicate that the combination of recurrent processing, convolutional feature extraction, and Gaussian modeling provides a more comprehensive analytical framework. The 2.7% accuracy improvement over GNN models demonstrates this advantage quantitatively.

Metaheuristic optimization has been previously applied to fake news detection, as exemplified by Yildirim[35] using Grey Wolf Optimization and Particle Swarm Optimization. However, our MFWO framework represents advancement through its integration of multiple optimization strategies and direct coupling with deep learning architectures. The ablation study results confirm this contribution, showing a 3.3% performance improvement attributable to MFWO optimization.

Analysis of linguistic and structural features

Examining the features identified by our model as most discriminative for fake news classification provides insights into the linguistic and structural characteristics of deceptive content. Through feature importance analysis based on attribution methods, we identified several consistent patterns:

1. **Lexical features**: Fake news articles frequently employed more extreme or emotionally charged language, with significantly higher frequencies of superlatives and emotionally intense adjectives ($p < 0.01$) compared to authentic news.

2. **Structural characteristics**: Authentic news typically exhibited more complex sentence structures, including higher rates of subordinate clauses and longer average sentence lengths (mean difference: 3.6 words per sentence, $p < 0.05$).
3. **Source and reference patterns**: Fake news demonstrated fewer explicit source attributions and external references per word count (mean difference: 0.82 references per 100 words, $p < 0.01$).
4. **Temporal indicators**: Fabricated content showed less specific temporal anchoring, with fewer precise dates and times mentioned ($p < 0.05$) and more general temporal references.
5. **Certainty markers**: Fake news exhibited higher frequencies of absolute certainty indicators (e.g., "definitely," "absolutely," "undoubtedly") compared to authentic reporting, which more frequently employed qualification and nuance ($p < 0.01$).

These patterns align with theoretical perspectives on deceptive communication, which suggest that fabricated content often exhibits linguistic distancing, excessive certainty, and emotional manipulation techniques[46]. The consistent identification of these patterns across multiple datasets supports their validity as generalizable indicators of potentially deceptive content.

Conclusion

This research addresses the critical challenge of fake news detection through a novel hybrid computational framework integrating deep learning architectures with metaheuristic optimization. Our methodology synthesizes recurrent LSTM networks for sequential feature extraction, Convolutional Gaussian Perceptron Neural Networks for probabilistic classification, and Metaheuristic Moth-Flame Whale Optimization for parameter refinement. The framework incorporates innovative lexicon-based scoring mechanisms to construct latent variables addressing missing data challenges in social media content, while extracting sentiment-based probability scores from Part-of-Speech elements. Experimental validation across multiple benchmark datasets demonstrates superior performance, achieving 98% accuracy with statistically significant improvements over state-of-the-art approaches ($p < 0.05$). Comprehensive ablation studies quantify individual component contributions, revealing CGPNN's primary impact (7.9% accuracy enhancement), followed by MFWO optimization (3.3%) and recurrent LSTM features (4.8%). Error analysis identifies specific content categories challenging automated detection, including satirical, technical, hybrid factual-misleading, cross-cultural, and breaking news materials. While this hybrid approach represents significant advancement in automated misinformation detection, ongoing challenges in interpretability, domain adaptation, and computational efficiency require continued investigation. Future research should emphasize multimodal analysis integration, temporal dynamics modeling, enhanced explainability mechanisms, and collaborative human-AI system development to address the evolving complexity of digital misinformation.

Limitations & future work

While demonstrating high effectiveness in fake news detection, the proposed framework is subject to several methodological, practical, and ethical limitations that highlight avenues for future improvement. Methodologically, relying on English-language and politically centered benchmark datasets limits cross-cultural generalizability, while the Gaussian perceptron component introduces computational complexity that can hinder scalability in high-throughput scenarios. Additionally, the model exhibits the typical interpretability challenges of deep learning architectures, resulting in opaque decision-making processes. Another significant issue is the computational cost. Including the Convolutional Gaussian Perceptron Neural Network (CGPNN), while beneficial for probabilistic feature modeling, introduces substantial computational complexity, making the framework challenging to deploy in real-time or large-scale scenarios. Furthermore, common to many deep learning models, the framework functions as a "black box," offering little insight into its decision-making process. We acknowledge that explainability was not the main scope of this work; however, a discussion on SHAP, LIME, and attention visualization has now been added as future work to enhance transparency and trust.

On the practical front, the framework demands substantial computational resources, requiring over 35 h of training on high-end GPUs and exhibiting latency issues in real-time applications, with 19.1 ms processing time per article being problematic given that most misinformation spreads within the first hour of posting. Moreover, the system demonstrated a 24.3% vulnerability to adversarial attacks and experienced a 10–15% drop in accuracy during cross-domain testing, underscoring the need for robust domain adaptation strategies. Ethical considerations include the necessity for data anonymization, IRB compliance, cultural bias mitigation, and the integration of human oversight to prevent misuse of the model in content filtering.

To overcome these challenges, future research should explore six key directions: multimodal analysis incorporating visual data, temporal modeling of narrative shifts, integration of explainable AI techniques, few-shot learning for domain transfer, adversarial training to strengthen model resilience, and the development of nuanced classification systems that extend beyond binary labels.

Data availability

The data used in this study are public datasets. The related dataset links are provided within the manuscript information files.

Received: 4 June 2025; Accepted: 20 October 2025
Published online: 24 November 2025

References

1. Zubiaga, A. & Ji, H. Tweet, but verify: epistemic study of information verification on Twitter. *Social Netw. Anal. Min.* **4**, 1–12 (2014).
2. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Economic Perspect.* **31** (2), 211–236 (2017).
3. Asfand-e-Yar, M., Hashir, Q., Tanvir, S. H., Khalil, W. & Computers Classifying Misinformation of User Credibility in Social Media Using Supervised Learning. *Mater. Continua*, **75**(2) (2023).
4. Silverman, C. & Alexander, L. How teens in the Balkans are duping Trump supporters with fake news. *Buzzfeed News*. **3**, 874–888 (2016).
5. Ismail, N. & Ardalan-Raikes, A. United against Hate: Lessons from the far-right Riots in England. *Justice, Power Resist.* 1–9 (2025).
6. Bowles, J., Croke, K., Larreguy, H., Liu, S. & Marshall, J. Sustaining exposure to fact-checks: misinformation discernment, media consumption, and its political implications. *Am. Polit. Sci. Rev.*, 1–24. (2023).
7. Saeed, A. & Al Solami, E. Computers Fake news detection using machine learning and deep learning methods. *Mater. Continua*, **77**(2) (2023).
8. Bozkurt, B. A Thesis Proposal Claim Inspector Framework: A Hybrid Approach to Data Annotation using Fact-Checked Claims and LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 215–224 (2024).
9. Ilyas, M. A. et al. Fake news detection on social media using ensemble methods. Comput. Mater. Continua **81** (3) (2024).
10. Gottfried, M. & Shearer, E. News use across social media platforms 2016, Pew Research Center, May 2016. [Online]. Available: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/
11. Thibault, C. et al. A Guide to Misinformation Detection Datasets. arXiv preprint arXiv:2411.05060. (2024).
12. Yang, F. et al. Xfake: Explainable fake news detector with visualizations. In *The world wide web conference*, 3600–3604 (2019).
13. Zhang, L. et al. Mitigating social hazards: Early detection of fake news via diffusion-guided propagation path generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2842–2851 (2024).
14. Jadhav, D., & Singh, J. Web information extraction and fake news detection in twitter using optimized hybrid bi-gated deep learning network. *Multimedia Tools Applic.* **84**(11), 9471–9504 (2025).
15. Balshetwar, S. V. & Rs, A. Fake news detection in social media based on sentiment analysis using classifier techniques. *Multimedia Tools Appl.* **82** (23), 35781–35811 (2023).
16. Li, Y. et al. A survey on truth discovery. *ACM SIGKDD Explorations Newsl.* **17**(2), 1–16. https://doi.org/10.1145/2897350.2897352 (2016).
17. Azam, U., Rizwan, H. & Karim, A. Exploring data augmentation strategies for hate speech detection in roman urdu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4523–4531 (2022).
18. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., Din, M. U. & M Machine learning based approaches for detecting COVID-19 using clinical text data. *Int. J. Inform. Technol.* **12**, 731–739 (2020).
19. Comito, C., Falcone, D. & Talia, D. Mining human mobility patterns from social geo-tagged data. *Pervasive Mob. Comput.* **33**, 91–107 (2016).
20. Ruchansky, N., Seo, S. & Liu, Y. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806 (2017).
21. Albalawi, R. M., Jamal, A. T., Khadidos, A. O. & Alhothali, A. M. Multimodal Arabic rumors detection. *IEEE Access.* **11**, 9716–9730 (2023).
22. Liu, Y. & Wu, Y. F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32, No. 1 (2018).
23. Bian, T. et al. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 01, pp. 549–556 (2020).
24. Lyu, Y. et al. Interpretable and effective reinforcement learning for attacking against graph-based rumor detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9 (IEEE, 2023).
25. Sharma, K. et al. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol. (TIST).* **10** (3), 1–42 (2019).
26. Hung, M. C., Chen, A. P. & Yu, W. T. AI-driven intraday trading: applying machine learning and market activity for enhanced decision support in financial markets. *IEEE Access.* **12**, 12953–12962 (2024).
27. Do, H. H., Prasad, P. W., Maag, A. & Alsadoon, A. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Syst. Appl.* **118**, 272–299 (2019).
28. Xu, F. et al. Hierarchical graph attention networks for multi-modal rumor detection on social media. *Neurocomputing* **569**, 127112 (2024).
29. Dixit, D. K., Bhagat, A. & Dangi, D. Fake news classification using a fuzzy convolutional recurrent neural network. *Computers Mater. Continua*, **71**(3) (2022).
30. Gorai, J. & Shaw, D. K. Semantic difference-based feature extraction technique for fake news detection. *J. Supercomputing*. **80** (15), 22631–22653 (2024).
31. Ramya, S. P. & Eswari, R. Attention-based deep learning models for detection of fake news in social networks. *Int. J. Cogn. Inf. Nat. Intell. (IJCINI).* **15** (4), 1–25 (2021).
32. Chen, Y., Conroy, N. J. & Rubin, V. L. Misleading online content: recognizing clickbait as false news. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, 15–19 (2015).
33. Zhong, N., Jiang, X. & Yao, Y. Computers from detection to explanation: integrating temporal and spatial features for rumor detection and explaining results using LLMs. *Mater. Continua* **82**(3) (2025).
34. Alshahrani, H. J. et al. I. Hunter prey optimization with hybrid deep learning for fake news detection on Arabic corpus. *Comput. Mater. Continua* (2023).
35. Yildrim, G. A novel hybrid multi-thread metaheurisitic approach for the fake news detection in social media. *Appl. Intell.* **53**(9), 11182–11202 (2023).
36. Nadeem, M., Fang, W., Xu, B., Mohtarami, M. & Glass, J. FAKTA: An automatic end-to-end fact checking system. In *Proc. Conf. North American Chapter of the Association for Computational Linguistics* (Human Language Technologies), Minneapolis, MN, USA, 78–83 (2019). https://doi.org/10.18653/v1/N19-4014
37. Liang, Z. Fake news detection based on multimodal inputs. *Comput. Mater. Continua*, **75**(2) (2023).
38. Guo, H., Cao, J., Zhang, Y., Guo, J. & Li, J. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 943–951 (2018).
39. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsl.* **19** (1), 22–36 (2017).
40. Bondielli, A. & Marcelloni, F. A survey on fake news and rumour detection techniques. *Inf. Sci.* **497**, 38–55 (2019).
41. Pérez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R. Automatic detection of fake news. *ArXiv Preprint arXiv* :170807104 (2017).
42. Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. Lying words: predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.* **29** (5), 665–675 (2003).
43. Kaliyar, R. K., Goswami, A. & Narang, P. FakeBERT: fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools Appl.* **80** (8), 11765–11788 (2021).

44. He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. arXiv 2020. arXiv preprint arXiv:2006.03654 (2006).
45. Jin, Z., Cao, J., Zhang, Y., Zhou, J. & Tian, Q. Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multimedia*. **19** (3), 598–608 (2016).
46. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. Fakenewsnet: A data repository with news content, social context, and Spatiotemporal information for studying fake news on social media. *Big Data*. **8** (3), 171–188 (2020).
47. Zhou, X. & Zafarani, R. A survey of fake news: fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)*. **53** (5), 1–40 (2020).
48. Chinnasamy, P., Ayyasamy, R.K., Subramaniam, S., Sangodiah, A., Iftikhar, A., & Kiran, A. Fake social media profile identification and report Using machine learning. *In IEEE 2024 International Conference on signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, 1–5 (2024).
49. Tahayna, B.M., Ayyasamy, R.K., & Akbar, R. Automatic sentiment annotation of idiomatic expressions for sentiment analysis task. *IEEE Access*. **10**, 122234–122242 (2024).
50. Chinnasamy, P., Ayyasamy, R.K., Madhukar, G., Karthik, G., Bhargav, T., & Nayak, D.Y.K. Comment analyzer by sentiment analysis through natural language processing. *In 2024 10th IEEE International Conference on Communication and Signal Processing (ICCSP)*. 1123–1128 (2024).
51. Tufchi, S., Yadav, A. & Ahmed, T. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *Int. J. Multimedia Inform. Retr.* **12** (2), 28 (2023).
52. Shan, F., Sun, H. & Wang, M. Multimodal social media fake news detection based on similarity inference and adversarial Networks. *Comput. Mater. Continua*, **79**(1) (2024).
53. Cao, J. et al. Fake news detection based on cross-modal ambiguity computation and multi-scale feature fusion. *Comput. Mater. Continua* **83** (2) (2025).
54. Li, Y., Dai, M., & Zhang, S. FHGraph: A Novel framework for fake news detection using graph contrastive learning and LLM. *Comput. Mater. Continua*, **83**(1) (2025).

Acknowledgements
The authors would like to express gratitude to Department of Computer Science, Mizan Tepi University, Mizan Teferi, Ethiopia.

Author contributions
Ramesh Kumar Ayyasamy: Conceptualization, Methodology, Coding, Validation, Writing – original draft, Formal Analysis, Visualization. Chinnasamy Ponnusamy: Conceptualization, Methodology, Supervision, Writing – original draft. Kovvuri N Bhargavi: Conceptualization, Data Curation, Validation, Resources, Writing review & editing. Srikanth Cherukuvada: Conceptualization, Methodology, Investigation, Writing – review & editing. G.Charles Babu: Validation, Methodology, Investigation, Writing – review & editing. S.Amutha: Validation, Methodology, Investigation, Writing – review & editing. Dawit Tadesse Gamu: Validation, Methodology, Investigation, Writing – review & editing.

Funding
The author(s) received no specific funding for this work.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to R.K.A. or D.T.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.