Insights in
Computer Science

https://acadpub.com/ics

Cuckoo Swarm Optimization: An Intelligent Feature Selection Approach for Enhanced Fake News Detection

[1]Nikita Garg, [1]Pritam Singh Negi, [2]Nischay Garg

[1]Department of Computer Science and Engineering, School of Engineering and Technology, H.N.B. Garhwal University, Srinagar Garhwal, Uttarakhand 246174, India
[2]Department of Computer Science and Engineering, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

About the Article

Open Access

Review Article

DOI: 10.68163/ics.2026.67

How to Cite:
N. Garg, P. S. Negi, and N. Garg, "Cuckoo swarm optimization: An intelligent feature selection approach for enhanced fake news detection," *Insights Comput. Sci.*, vol. 2, pp. 18–29, 2026.

Keywords:
Cuckoo search, fake news detection, feature selection, machine learning

Corresponding author:
Nikita Garg,
Department of Computer Science and Engineering, School of Engineering and Technology, H.N.B. Garhwal University, Srinagar Garhwal, Uttarakhand 246174, India

ABSTRACT

Objective: This study aimed to develop and evaluate a machine learning-based framework for text-based fake news detection by integrating Cuckoo Search (CS) algorithm-based feature selection with conventional machine learning classifiers.
Materials and Methods: Five machine learning classifiers, namely Decision Tree, Support Vector Machine (SVM), XGBoost, Logistic Regression and Random Forest, were evaluated using a CS-based feature selection approach. Model performance was assessed using accuracy, precision, recall and F1-score.
Results: The Random Forest classifier combined with CS-based feature selection demonstrated the highest overall performance, achieving an accuracy of 91%, precision of 90%, recall of 92% and F1-score of 91%. Its performance exceeded that of the other evaluated classifiers under the investigated conditions. In contrast, the Decision Tree classifier exhibited comparatively lower performance, which may be attributable to its greater sensitivity to the structure and characteristics of the selected feature space.
Conclusion: The findings demonstrate that integrating metaheuristic feature selection with conventional machine learning algorithms can provide an effective and computationally manageable approach for text-based fake news detection. In particular, CS-based feature selection combined with Random Forest effectively identified a discriminatory feature subset and improved classification performance, highlighting its potential for automated misinformation detection on digital platforms.

INTRODUCTION

The rapid proliferation of fake news across digital platforms has emerged as a critical societal challenge, contributing to the dissemination of misinformation, public confusion and erosion of trust. The expansion of social media and online news platforms has made the manual identification of fake news increasingly difficult because of the volume, velocity and diversity of information generated and shared online. Inaccurate or misleading information can have serious consequences, including influencing public opinion and contributing to social and political instability. Consequently, there is an increasing need for automated approaches capable of detecting misleading information efficiently and at scale. Fake news detection has therefore attracted considerable research attention, with machine learning (ML) and artificial intelligence (AI) providing important computational approaches to address this challenge [1], [2], [3], [4].

Traditional manual approaches to fake news detection are often inefficient and time-consuming because they rely heavily on human judgment, domain expertise and expert verification. In contrast, automated ML-based systems can process large volumes of textual information and provide scalable solutions for classification. Various ML algorithms, including Decision Trees, Random Forest, Support Vector Machines (SVMs) and boosting-based methods, have been investigated for fake news detection [5], [6], [7], [8], [9]. Deep learning approaches have also been

© The Authors, 2026. Published by the Academia Publications. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (CC-BY) 4.0. (http://creativecommons.org/licenses/by/4.0)

ACADEMIA Publications

explored, particularly because of their ability to learn complex textual representations and contextual relationships [10], [11], [12]. However, the performance of these approaches can be substantially influenced by the quality, relevance and dimensionality of the features provided to the classifier. Therefore, effective feature representation and feature selection remain important components of fake news detection systems [3], [4], [13].

Feature selection aims to identify informative features while eliminating irrelevant or redundant variables, thereby reducing the dimensionality of the feature space and potentially improving computational efficiency and classification performance. Metaheuristic optimization techniques offer an alternative strategy for searching large and complex feature spaces. Population-based optimization methods have been applied to feature-selection problems, including Genetic Algorithms and other evolutionary optimization techniques [14], [15], [16]. Among these methods, Cuckoo Search (CS) is a population-based metaheuristic optimization algorithm introduced by Yang and Deb, which employs Lévy-flight-based exploration to search for promising solutions [17]. This search mechanism makes CS suitable for optimization problems involving a large number of possible feature combinations. In the context of fake news detection, such an optimization strategy can be employed to identify informative subsets of textual features before classification, potentially reducing feature redundancy while retaining discriminative information.

Text representation constitutes another important component of a text-based fake news detection framework. Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used representation technique in information retrieval and natural language processing that assigns weights to terms according to their importance within individual documents and across the corpus [18]. By transforming textual information into numerical feature vectors, TF-IDF provides a suitable representation for conventional ML classifiers. When combined with feature-selection techniques, TF-IDF representations can potentially be reduced to a more informative subset of features, thereby minimizing the influence of irrelevant or redundant terms and improving the efficiency of subsequent classification.

The present study investigates the application of CS for feature selection in text-based fake news detection. The dataset comprises 6,335 records with four columns: S. No., Title, Text and Label. The Title and Text columns contain the textual information used for fake news classification, whereas the Label column categorizes the news as either "fake" or "real". In this study, the Title and Text fields are used for textual feature extraction and classification. The preprocessing stage includes tokenization, stopword removal and text normalization. Subsequently, 1,000 TF-IDF features are extracted from the processed textual data. These features are then subjected to CS-based feature selection to identify a potentially more informative feature subset. The selected features are subsequently evaluated using five ML classifiers, namely Decision Tree, Support Vector Machine (SVM), XGBoost, Logistic Regression and Random Forest.

Random Forest is an ensemble learning method that combines predictions from multiple decision trees and has demonstrated robustness across a range of classification tasks [5]. In the present study, Random Forest is evaluated alongside the other classifiers to determine its effectiveness when combined with the feature representation optimized using CS. The results indicate that the Random Forest classifier achieved the highest performance under the investigated conditions, with an accuracy of 91%, precision of 90%, recall of 92% and F1-score of 91%. These findings suggest that integrating metaheuristic feature selection with conventional ML algorithms may provide an effective approach for text-based fake news classification.

The principal contribution of this study is the integration of CS-based feature selection with conventional ML classifiers for fake news detection. Rather than focusing exclusively on classifier selection, the proposed framework incorporates optimization of the textual feature space before classification. This approach provides a computationally manageable and potentially interpretable framework for the investigated text-based classification setting. Nevertheless, fake news detection represents a broader research problem involving textual, social, contextual and multimodal information. Recent studies have emphasized the importance of multimodal information and social context in detecting increasingly complex forms of misinformation [19], [20], [21], [22], [23], [24], [25]. Therefore, the proposed framework may provide a foundation for future research aimed at developing more generalizable and context-aware fake news detection systems.

The remainder of this paper is organized as follows. Section 2 presents a review of the existing literature and discusses relevant developments in fake news detection. Section 3 describes the proposed model and methodology. Section 4 presents the results and their analysis. Section 5 discusses the findings in relation to existing research. Finally, Section 6 presents the conclusions and future scope of the study.

Related work: The detection of fake news has attracted considerable research attention with the rapid expansion of social media and online information platforms. Existing studies have investigated a broad range of approaches, including traditional machine learning, deep learning, ensemble learning, social-context analysis and multimodal methods, for identifying misleading information. Early research established fake news detection as a data mining and classification problem, emphasizing the importance of

textual characteristics and computational methods for distinguishing false information from reliable content [1]. Subsequent studies demonstrated that fake news detection cannot be addressed solely through textual analysis, as social context, propagation behaviour and other information associated with online content can provide additional discriminative signals [2], [20], [23].

Traditional machine learning methods remain important components of fake news detection frameworks. Algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM) and gradient-boosting methods can classify textual representations following appropriate preprocessing and feature extraction [5], [6], [7], [8], [9]. Random Forest, in particular, is an ensemble learning algorithm that combines the predictions of multiple decision trees, providing comparatively robust performance in classification tasks involving high-dimensional feature spaces [5]. The SVM represents another established classification framework that is particularly suitable for high-dimensional feature spaces [6], whereas XGBoost provides a scalable tree-boosting approach capable of modelling complex relationships among features [7]. Comparative studies have indicated that classifier performance can vary depending on the dataset, feature representation and specific characteristics of the fake news detection task [3], [13], [17]. These findings highlight the importance of evaluating classifiers within the context of the underlying data and feature representation rather than assuming that a particular algorithm will consistently provide optimal performance.

Feature representation is therefore a critical component of text-based fake news detection. Term Frequency-Inverse Document Frequency (TF-IDF) provides a conventional and computationally efficient approach for transforming textual information into numerical representations by assigning weights to terms according to their importance within individual documents and across the corpus [18]. However, textual datasets can generate a large number of features, many of which may be redundant, irrelevant, or relatively uninformative for classification. Consequently, feature selection has been investigated as a strategy for reducing the dimensionality of the feature space while retaining features that contribute meaningfully to classification performance. Recent studies have specifically examined optimization-based feature selection for fake news detection, including approaches based on Genetic Algorithms and other population-based optimization techniques [14], [15], [16].

Metaheuristic optimization methods are particularly relevant to feature-selection problems because the number of possible feature combinations can increase exponentially with the dimensionality of the feature space. Cuckoo Search (CS), introduced by Yang and Deb, is a population-based optimization algorithm that employs Lévy-flight-based exploration to search for promising solutions [17]. This search mechanism provides a theoretical basis for applying CS to feature-selection problems, in which candidate solutions can represent different subsets of the available features. Previous studies on CS-based feature selection have demonstrated its applicability to classification-oriented feature-selection problems [14], [15]. Recent research has also investigated Genetic Algorithm-based feature selection specifically for fake news detection [16]. Collectively, these studies provide a rationale for investigating metaheuristic feature-selection methods as alternatives or complements to conventional feature-ranking approaches.

Deep learning has further expanded the methodological scope of fake news detection. Studies such as Kumar et al. [10] have investigated deep learning models for automatically learning representations from news data whereas Kaliyar et al. [11] proposed FakeBERT, demonstrating the application of BERT-based contextual representations to fake news detection. FNDNet also investigated a deep convolutional neural network architecture for detecting fake news [12]. These approaches demonstrate the ability of deep learning models to learn complex and context-sensitive representations from textual data. However, such models may require greater computational resources and can be more difficult to interpret than conventional feature-based machine learning approaches. Therefore, computationally efficient approaches that can effectively exploit informative textual features remain relevant, particularly in settings where computational resources or model interpretability are important considerations.

Research has also increasingly focused on multimodal fake news detection. Multimodal approaches seek to integrate textual information with visual and other complementary sources of information because deceptive content can involve multiple interacting information modalities. EANN, for example, investigated multimodal fake news detection using an event-adversarial neural network framework [21], whereas MVAE explored multimodal representation learning for fake news detection [22]. Other studies have examined multimodal information and social-context signals to address the limitations associated with relying exclusively on textual content [24], [25]. Recent surveys have further emphasized the increasing importance of multimodal approaches for detecting complex forms of online misinformation [19], [25]. These developments indicate that incorporating multiple information sources can potentially enhance the ability of detection systems to identify misinformation that cannot be reliably characterized through textual features alone.

Explainability and social context have also emerged as important research directions in fake news detection. The dEFEND framework incorporated explainable mechanisms into fake news detection and demonstrated the value of

considering both news content and associated user comments [26]. Research focusing on social context has similarly shown that information beyond the textual content of a news article can provide useful signals for identifying misinformation [20], [23]. These findings suggest that effective future fake news detection systems should consider not only linguistic characteristics but also contextual, social and propagation-related information.

Several benchmark datasets have contributed substantially to the development and evaluation of fake news detection systems. The LIAR dataset, introduced by Wang, provides a benchmark containing labelled statements for fake news classification [27]. Other datasets have been developed for specific misinformation scenarios, including COVID-19-related fake news [28]. Such datasets have enabled researchers to evaluate and compare different feature representations, machine learning algorithms and deep learning approaches under varying data characteristics. However, differences in dataset composition, class distributions, textual characteristics and annotation strategies can substantially influence model performance and should therefore be considered when interpreting comparative results.

Recent surveys and comparative studies have further emphasized the importance of considering the interaction among dataset characteristics, feature representation, feature selection and classification algorithms. Alghamdi et al. [13] provided a comprehensive review of machine learning approaches for fake news detection, whereas Farhangian et al. [17] presented a taxonomy and comparative analysis of fake news detection methods. Earlier reviews similarly emphasized that fake news detection is a multidisciplinary problem involving content analysis, machine learning, social-context information and other computational signals [3], [4]. Collectively, these findings indicate that no single classifier or feature representation is universally optimal and that the effectiveness of a detection framework depends substantially on the characteristics of the available data and the modelling strategy adopted. Based on the existing literature, there remains a need to investigate approaches that combine computationally efficient textual representations with intelligent feature-selection mechanisms and conventional machine learning classifiers. The present study addresses this research area by investigating CS-based feature selection applied to TF-IDF representations and evaluating the resulting feature subset using Decision Tree, Support Vector Machine, XGBoost, Logistic Regression and Random Forest classifiers. The proposed approach focuses on reducing the dimensionality of the textual feature space before classification rather than relying exclusively on increasingly complex deep learning architectures. This strategy provides a theoretically motivated and computationally manageable framework for text-based fake news detection, while also providing a foundation for future extensions incorporating social context, propagation behaviour and multimodal information.

Proposed work: In this study, a text-based fake news detection framework is proposed that integrates Cuckoo Search (CS)-based feature selection with conventional machine learning classifiers. The overall framework comprises data preprocessing, Term Frequency-Inverse Document Frequency (TF-IDF)-based feature extraction, dataset partitioning, CS-based feature selection, classification and performance evaluation. Initially, the news dataset is preprocessed to remove irrelevant textual elements, including punctuation, stopwords and other unnecessary content. The preprocessed title and text data are subsequently transformed into numerical feature representations using the TF-IDF technique. A total of 1,000 TF-IDF features are extracted to represent the textual information. Following feature extraction, the dataset is partitioned into training and testing subsets using an 80:20 ratio. CS is then applied to the training feature space to identify an informative subset of features. The selected feature subset is subsequently used to train five machine learning classifiers, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost and Decision Tree. The trained classifiers are evaluated using the corresponding test data based on accuracy, precision, recall and F1-score. The complete workflow of the proposed framework, encompassing data preprocessing, feature extraction, feature selection, classification and performance evaluation, is illustrated in Fig. 1.

Figure 1 illustrates the proposed framework developed in this study, which integrates Cuckoo Search (CS)-based feature selection with five machine learning algorithms for fake news detection. The workflow begins with data preprocessing, followed by feature extraction from the processed textual data. The extracted features are subsequently optimized using the CS algorithm to identify a more informative and relevant feature subset. The selected features are then used to train and evaluate five machine learning classifiers, enabling a comprehensive assessment of their classification performance. Overall, the proposed framework is designed to enhance fake news detection performance while reducing feature dimensionality and improving computational efficiency.

MATERIALS AND METHODS

Dataset collection: A publicly available news dataset obtained from Kaggle was used in this study. The dataset comprises 6,335 records and four columns: S.No., Title, Text and Label. The Title and Text fields contain the textual information used for fake news classification, whereas the Label field represents the corresponding class, namely "fake" or "real" [27]. The analysis focused on the Title and

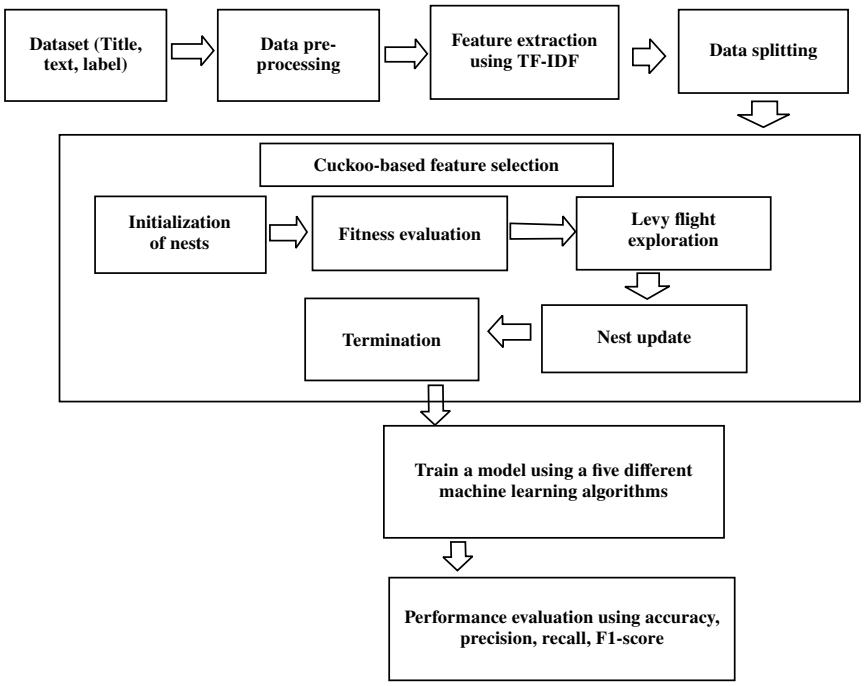

Fig. 1: Proposed model

Text fields because they contain the primary textual information required for feature extraction and subsequent classification.

Data preprocessing: Following data collection, the textual data were preprocessed to transform the raw information into a suitable format for feature extraction. The preprocessing procedure included the removal of punctuation marks, stopwords and other unnecessary textual elements that could introduce noise into the analysis. The text was also converted to lowercase to ensure consistency and minimize the effects of differences in letter case. These preprocessing steps generated a more standardized textual representation and provided suitable input for the subsequent TF-IDF feature-extraction procedure.

Feature extraction: Following preprocessing, the textual information was transformed into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The TF-IDF assigns weights to terms according to their frequency within an individual document and their relative importance across the corpus [18]. In this study, 1,000 TF-IDF features were extracted to represent the textual information. The Title and Text fields were used for feature extraction, resulting in a TF-IDF matrix with dimensions of (6335, 1000), where each row represents a news record and each column represents a textual feature. This numerical representation enabled the textual information to be subsequently processed by the Cuckoo Search-based feature-selection procedure.

Data splitting: Following TF-IDF feature extraction, the dataset was divided into training and testing subsets using an 80:20 ratio. The training subset was used for feature selection and model training, whereas the testing subset was reserved for evaluating model performance on unseen data. Dataset partitioning was performed using the train_test_split function from the scikit-learn framework, with random_state = 42 to ensure reproducibility. To minimize the risk of information leakage, the Cuckoo Search feature-selection procedure was applied exclusively to the training data. The selected feature subset was subsequently applied to both the training and testing representations before classifier evaluation.

Feature selection: The Cuckoo Search Algorithm was employed for feature selection in the fake news detection framework. Cuckoo Search is a population-based metaheuristic optimization technique inspired by the brood-parasitic behaviour of cuckoo birds and incorporates Lévy-flight-based exploration [17]. In the proposed framework, each candidate solution represents a potential subset of the available TF-IDF features. The objective of the optimization process was to identify a feature subset that provides effective classification performance while reducing the dimensionality of the original feature space.

Cuckoo search algorithm: The Cuckoo Search optimization procedure comprised the following stages:

- **Initialization:** A population of nests was randomly initialized, with each nest representing a candidate feature subset

- **Fitness evaluation:** The fitness of each candidate feature subset was evaluated using the training data and a Random Forest classifier.
- **Lévy-flight exploration:** Lévy-flight-based movements were employed to generate new candidate solutions and explore the feature-search space
- **Nest update:** Each newly generated candidate solution was compared with the corresponding existing solution. If the new solution provided better fitness, it replaced the existing solution
- **Best solution selection:** Following the specified number of iterations, the candidate solution with the highest fitness value was selected as the final feature subset

Feature-selection process: The feature-selection procedure consisted of the following steps:

- **Random initialization:** Five nests were initialized, with each nest representing a candidate subset of the available features.
- **Fitness evaluation:** The fitness of each nest was calculated based on the classification performance obtained using the corresponding training feature subset
- **Lévy-flight exploration:** Lévy-flight-based exploration was employed to generate new candidate feature subsets
- **Nest update:** If a newly generated candidate solution produced a higher fitness value than the existing solution, it replaced the corresponding nest.
- **Termination:** The optimization process was performed for 10 iterations. Upon completion, the nest with the highest fitness value was selected
- **Feature selection:** Features corresponding to values greater than 0.5 in the selected solution vector were retained

Pseudocode: The pseudocode for the Cuckoo Search-based feature-selection procedure is presented below:

```
    Initialize 5 candidate nests
    Evaluate the fitness of each nest using the training data
Select the nest with the highest fitness
For 10 iterations:
    For each nest:
    Generate a new candidate solution using Lévy flight
    Restrict candidate values to the range [0, 1]
    Evaluate the new candidate using the training data
    If new fitness > current fitness:
    Replace the current nest
    Update its fitness
Select the nest with the highest fitness
Select features whose final values exceed 0.5
```

Execution conditions: The feature-selection procedure was configured with five nests and a maximum of 10 iterations. Each nest represented a candidate feature subset and the optimization process iteratively evaluated and updated these candidate solutions according to their fitness values. Upon completion of the optimization process, the nest with the highest fitness value was selected as the optimal solution. Features corresponding to values greater than 0.5 in the selected solution vector were retained. Under the investigated conditions, the final selected feature subset comprised 502 features from the original 1,000 TF-IDF features.

Machine learning algorithms: The selected feature subset was evaluated using five machine learning classifiers, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost and Decision Tree.

Logistic regression: Logistic Regression is a supervised classification algorithm commonly used for binary classification problems. It models the relationship between input features and the probability that an observation belongs to a particular class. In this study, Logistic Regression was employed to distinguish between fake and real news based on the selected textual features.

Random forest: Random Forest is an ensemble learning method that combines multiple decision trees to generate a final classification. By aggregating predictions from multiple trees, Random Forest can provide greater robustness than an individual decision tree [7]. In the present study, Random Forest was also employed as the fitness classifier during the Cuckoo Search-based feature-selection process.

Support vector machine: Support Vector Machine (SVM) is a supervised learning algorithm designed to identify an optimal decision boundary between classes. It is particularly suitable for classification problems involving high-dimensional feature spaces. The SVM identifies a separating hyperplane while maximizing the margin between classes [23]. In this study, SVM was used to classify news articles as fake or real based on the selected textual features.

XGBoost: XGBoost is a scalable gradient-boosting framework based on an ensemble of decision trees. It sequentially constructs decision trees to reduce the errors of preceding models and incorporates regularization mechanisms to control model complexity [24]. In this study, XGBoost was employed as one of the comparative classifiers for distinguishing between fake and real news.

Decision tree: Decision Tree is a supervised learning algorithm that recursively partitions the feature space into

subsets according to feature values. Its hierarchical tree structure makes the resulting classification process relatively straightforward to interpret. However, an individual decision tree may be more susceptible to overfitting than ensemble approaches such as Random Forest [14]. In this study, Decision Tree was evaluated as one of the comparative classifiers using the selected textual features.

Performance evaluation: The performance of the five classifiers was evaluated using accuracy, precision, recall and F1-score. These metrics were derived from the confusion matrix, which comprises four components: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These components provide the basis for calculating several widely used classification performance metrics, including accuracy, precision, recall and F1-score.

Accuracy: Accuracy represents the overall proportion of correctly classified instances, including both positive and negative predictions, among all evaluated instances. It is calculated as the ratio of the total number of correct predictions to the total number of predictions. However, accuracy may provide a misleading assessment when the dataset contains substantially imbalanced classes.

Precision: Precision measures the proportion of correctly identified positive instances among all instances predicted as positive. It is calculated as the ratio of true positives to the sum of true positives and false positives and provides an indication of the model's ability to minimize false-positive predictions.

Recall: Recall measures the ability of a classifier to correctly identify relevant positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives and indicates the effectiveness of the model in detecting positive instances.

F1-score: The F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification performance when both false-positive and false-negative errors are important. It is particularly useful when class distributions are imbalanced or when a balance between precision and recall is required.

RESULTS

In this study, a news dataset comprising the columns S. No., Title, Text and Label was analyzed. The Title and Text fields were used for textual preprocessing and feature extraction. Following preprocessing, including tokenization, stopword removal and text normalization, 1,000 TF-IDF features were extracted. Cuckoo Search was subsequently applied as a feature-selection mechanism to identify a reduced subset of informative features. As described in the methodology, 502 features were retained following the feature-selection procedure. The selected features were then evaluated using five machine learning classifiers, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost and Decision Tree. Model performance was assessed using accuracy, precision, recall and F1-score and the results are summarized in Table 1. A comparative evaluation of the accuracy is presented in Table 1 whereas Table 2 presents a comparative evaluation of the precision, recall and F1-score achieved by the five machine learning algorithms investigated in this study.

Table 1 shows that Random Forest achieved the highest accuracy of 91%, followed by SVM and XGBoost, both of which achieved an accuracy of 90%. Logistic Regression achieved an accuracy of 88%, whereas Decision Tree exhibited the lowest accuracy, at 78%.

As shown in Table 2, Random Forest achieved a precision of 90%, recall of 92% and F1-score of 91%. Although SVM achieved a slightly higher precision of 91%, its recall and F1-score were lower than those of Random Forest. XGBoost achieved a precision of 91%, recall of 91% and F1-score of 91%, demonstrating performance comparable to Random Forest across these evaluation metrics. Overall, Random Forest achieved the highest accuracy and recall among the evaluated classifiers, while its F1-score was comparable to that of XGBoost.

The relatively high recall achieved by Random Forest indicates that the classifier correctly identified a large proportion of the relevant fake-news instances in the

Table 1: Model accuracy comparison

Model	Accuracy (%)
Logistic Regression	88
Random Forest	91
Support Vector Machine	90
XGBoost	90
Decision Tree	78

Table 2: Precision, recall, F1-score for each model

Model	Precision (%)	Recall (%)	F1-Score (%)
Logistic regression	90	87	88
Random forest	90	92	91
Support vector machine	91	90	90
XGBoost	91	91	91
Decision tree	78	79	78

evaluated dataset. Its balanced precision, recall and F1-score further suggest relatively consistent classification performance across the evaluated metrics. In contrast, Decision Tree exhibited the lowest overall performance, with an accuracy of 78%, precision of 78%, recall of 79% and F1-score of 78%. The comparatively lower performance of a single Decision Tree relative to Random Forest is theoretically consistent with the ensemble architecture of Random Forest, in which multiple decision trees are combined to improve model robustness. Nevertheless, this observed difference should be interpreted in the context of the investigated dataset and experimental setting rather than as evidence that Decision Trees are universally inferior to ensemble-based approaches.

Logistic Regression achieved an accuracy of 88%, whereas both SVM and XGBoost achieved accuracies of 90%. These findings indicate that the selected feature subset contained discriminative information that could be effectively utilized by different classification algorithms. The relatively small performance differences among Random Forest, SVM and XGBoost further suggest that classifier performance is influenced by the interaction between the selected feature representation and the learning algorithm.

Overall, the findings demonstrate the potential effectiveness of integrating Cuckoo Search-based feature selection with conventional machine learning classifiers for the investigated text-based fake news detection task. Random Forest achieved the highest accuracy of 91% and recall of 92%, whereas XGBoost achieved a comparable F1-score of 91%. These results support the potential utility of intelligent feature selection for reducing the dimensionality of the original textual feature space while retaining features that contribute meaningfully to classification performance.

DISCUSSION

The findings of the present study indicate that feature selection plays an important role in enhancing the effectiveness of machine learning-based fake news detection. In the proposed framework, Cuckoo Search (CS) was employed to identify a reduced subset of informative TF-IDF features before classification. Random Forest demonstrated the highest overall performance, achieving an accuracy of 91%, precision of 90%, recall of 92% and F1-score of 91%. These findings suggest that combining dimensionality reduction through metaheuristic feature selection with an ensemble-based classifier can provide an effective strategy for text-based fake news detection. The observed performance is consistent with the broader literature, which identifies feature representation, feature selection and classifier selection as important components of automated fake news detection systems [3], [4], [13], [17].

Recent reviews have demonstrated that traditional machine learning approaches remain relevant to fake news detection, particularly when textual information is represented using carefully selected features. Alghamdi et al. [13] emphasized that machine learning-based fake news detection involves multiple interdependent stages, including data preprocessing, feature representation, feature selection and classification. Similarly, Farhangian et al. [17] reported that the effectiveness of fake news detection models depends substantially on dataset characteristics, feature representation and the selected classification strategy. Previous surveys have likewise highlighted that detection performance is strongly influenced by the interaction among the available information, feature representation, learning algorithm and characteristics of the target dataset [3], [4]. Therefore, the performance observed in the present study supports the view that optimizing the representation of textual information can be as important as selecting an appropriate classifier.

A possible explanation for the strong performance of Random Forest is its ensemble architecture. Random Forest constructs multiple decision trees using randomized subsets of observations and features and aggregates their predictions to obtain the final classification. This mechanism can enhance robustness to noise and reduce the sensitivity associated with an individual decision tree [5]. In the present framework, CS reduces the original feature space by retaining a subset of potentially informative features. The subsequent application of Random Forest enables the classifier to exploit nonlinear relationships among the selected textual features while benefiting from the robustness of ensemble learning. The combination may therefore reduce the influence of irrelevant or redundant TF-IDF features and facilitate more stable classification of fake and real news.

The comparatively lower performance of Decision Tree relative to Random Forest also has a plausible methodological explanation. An individual decision tree can be sensitive to the specific structure of the training data and may generate complex decision rules when the feature space contains numerous textual variables. Such characteristics can increase the risk of overfitting. Random Forest addresses this limitation by aggregating multiple randomized trees, thereby improving robustness and reducing dependence on the structure of any single tree [5], [8]. The substantially higher performance of Random Forest compared with the individual Decision Tree in the present study is therefore theoretically consistent with the established advantages of ensemble learning. Nevertheless, this observation should be interpreted within the specific dataset and experimental conditions investigated rather than as evidence that Decision Trees are universally inferior to ensemble methods.

The results obtained with Logistic Regression, Support Vector Machine (SVM) and XGBoost further demonstrate that classifier performance can vary according to the selected textual representation. Logistic Regression achieved an accuracy of 88%, whereas SVM and XGBoost achieved accuracies of 90%. These findings indicate that the selected feature subset contained discriminative information that could be exploited by both linear and nonlinear classification approaches. SVM provides an established discriminative learning framework for high-dimensional feature spaces, whereas XGBoost provides a scalable tree-boosting approach for modelling complex feature relationships [6], [7]. The slightly higher performance of Random Forest suggests that the selected feature representation may contain interactions or nonlinear patterns that are effectively captured by an ensemble of decision trees. This observation is consistent with comparative research indicating that no single feature representation or classifier is universally optimal across fake news datasets, with performance instead depending on the interaction among the representation, classifier and dataset characteristics [17], [9].

The role of CS in the proposed framework is particularly relevant from an optimization perspective. CS is a population-based metaheuristic optimization technique introduced by Yang and Deb that incorporates Lévy-flight-based exploration to search a solution space [17]. In feature-selection problems, a candidate solution can be represented as a subset of the available features, thereby transforming feature selection into a search problem involving numerous possible feature combinations. An important advantage of this strategy is that the optimization process can explore combinations of textual features rather than relying exclusively on individual feature-ranking criteria. Consequently, the feature-selection stage may eliminate less informative variables while retaining feature combinations that contribute to classification performance. Previous studies on optimization-based feature selection have demonstrated the potential of population-based search techniques for identifying informative feature subsets in classification problems [14], [15]. These findings provide theoretical support for the application of metaheuristic optimization as a feature-selection mechanism in text-based classification.

The present findings also complement earlier research emphasizing the importance of combining textual characteristics with appropriate computational approaches for fake news detection. Shu et al. [1] highlighted the complexity of fake news detection, particularly because misleading information may be intentionally constructed to appear credible and textual content alone may not always provide sufficient evidence for reliable identification. Similarly, Vosoughi et al. [2] demonstrated the substantial impact of false information in online environments and reported differences in the propagation characteristics of false and true news. These observations reinforce the need for automated detection approaches capable of efficiently processing large volumes of textual information. Research incorporating social context has further demonstrated that information beyond the news content itself can provide additional signals for fake news detection [20], [23]. The proposed CS-Random Forest framework contributes to this line of research by focusing on the optimization of textual features before classification rather than relying solely on classifier selection.

However, the present findings should not be interpreted as evidence that a text-based feature-selection approach can address all forms of fake news. Recent research has increasingly emphasized that misinformation can involve multiple modalities, including text, images, audio, video and social-context information. Reviews of multimodal disinformation detection indicate that integrating information from different modalities is an important research direction because textual content alone may not capture all relevant deceptive signals [19], [25]. Multimodal approaches have demonstrated the potential value of combining heterogeneous information sources when addressing complex forms of online misinformation [21], [22], [24]. Therefore, although the present results support the utility of optimized textual feature selection, they also highlight the need for future extensions incorporating contextual and multimodal information.

Generalizability represents another important consideration. The present study employed a publicly available dataset and focused exclusively on textual information represented using TF-IDF features. Consequently, the reported performance should be interpreted within the characteristics of the investigated dataset rather than considered a universal performance level for fake news detection across different environments. Comparative research has shown that the optimal feature representation and classification strategy can vary among datasets and that context-dependent representations may provide advantages in more complex settings [3], [4], [17]. Research examining social context has likewise demonstrated that information associated with users, publishers and online interactions can provide additional signals for identifying fake news [20], [23]. Accordingly, the proposed framework should be regarded as a computationally interpretable and potentially effective approach for the investigated text-based setting rather than as a universally optimal solution.

The findings also illustrate the potential value of combining classical machine learning with computationally efficient feature optimization. Contemporary fake news detection research increasingly employs deep learning, multimodal models and contextual representations because of their capacity to learn complex semantic and contextual

patterns [10], [11], [12], [19], [25]. Nevertheless, conventional approaches based on TF-IDF and machine learning remain relevant when computational efficiency, interpretability and relatively lower resource requirements are important considerations [3], [13]. The application of CS for feature optimization provides an intermediate strategy in which the dimensionality of the feature space is intelligently reduced before conventional classifiers are applied. Thus, the contribution of the present study is not to replace advanced neural approaches but rather to demonstrate the continued potential of optimized feature-selection strategies for text-based fake news classification.

Overall, the findings support the central premise of this study that intelligent feature selection can complement conventional machine learning approaches for fake news detection. The accuracy of 91% and the balanced precision, recall and F1-score achieved by Random Forest indicate that the selected feature subset retained useful discriminatory information. These findings are broadly consistent with the literature emphasizing the importance of feature representation and classifier selection, while also highlighting the potential value of metaheuristic optimization as a feature-selection mechanism [13], [14], [15], [17]. At the same time, the broader literature indicates that future fake news detection systems will need to address multimodal information, social context, domain shifts, evolving misinformation patterns and explainability [19], [20], [21], [25], [26]. These considerations provide a basis for extending the proposed framework toward more generalizable, context-aware and robust fake news detection systems.

CONCLUSION AND FUTURE SCOPE

This study demonstrates the potential of intelligent feature selection for enhancing text-based fake news detection using conventional machine learning techniques. The proposed framework integrates Term Frequency-Inverse Document Frequency (TF-IDF)-based textual representation with Cuckoo Search-based feature selection, followed by the evaluation of multiple machine learning classifiers. Among the evaluated models, Random Forest demonstrated the best overall performance, achieving an accuracy of 91%, precision of 90%, recall of 92% and F1-score of 91%. These findings indicate that the selected feature subset retained useful discriminatory information and that an ensemble-based classifier could effectively exploit the optimized textual representation.

The results further demonstrate that classification performance varies according to the characteristics of the selected feature representation and the learning algorithm. Random Forest outperformed Decision Tree, Support Vector Machine (SVM), XGBoost and Logistic Regression under the investigated conditions. The comparatively lower performance of the individual Decision Tree may be associated with its greater sensitivity to the structure of the training data, whereas the ensemble architecture of Random Forest can enhance robustness by aggregating predictions from multiple decision trees. Nevertheless, these findings should be interpreted within the context of the investigated dataset and experimental setting rather than as evidence that Random Forest is universally optimal for fake news detection.

The application of Cuckoo Search provides an optimization-oriented strategy for feature selection by searching for informative combinations of textual features rather than relying exclusively on individual feature-ranking criteria. This approach can potentially reduce the influence of irrelevant and redundant features while retaining those that contribute to classification performance. Accordingly, the principal contribution of this study lies in integrating metaheuristic feature selection with conventional machine learning to develop a computationally manageable and relatively interpretable framework for text-based fake news detection.

Despite the encouraging findings, several limitations should be acknowledged. The proposed framework primarily relies on textual information represented using TF-IDF features and therefore does not capture the full range of contextual, social, or multimodal characteristics associated with misinformation. Furthermore, the reported performance is specific to the characteristics of the investigated dataset and should not be interpreted as a universal measure of model effectiveness across different domains, languages, datasets, or social media environments. Variations in language, news topics, misinformation strategies and data distributions may substantially affect the performance of the selected feature representation and classification algorithms.

Future research should investigate alternative and hybrid feature-selection strategies, including Genetic Algorithms, Particle Swarm Optimization and other population-based metaheuristic techniques. Further studies could also examine contextual and deep learning-based representations, including transformer-based models, as complementary or alternative approaches to conventional TF-IDF representations. Integration of textual, visual, social-context and other multimodal information represents another important direction for developing more comprehensive fake news detection systems. Cross-domain and multilingual evaluations would also be valuable for assessing the generalizability and robustness of the proposed framework across diverse datasets and linguistic settings. In addition, real-time detection, model explainability and adaptation to evolving misinformation patterns warrant further investigation.

Overall, the findings support the premise that intelligent feature selection can effectively complement conventional machine learning approaches for text-based fake news

detection. The combination of Cuckoo Search, TF-IDF and Random Forest provides a computationally manageable and theoretically motivated framework that can retain informative textual features while reducing the dimensionality of the feature space. The identified limitations and proposed research directions provide opportunities to extend this framework toward more generalizable, context-aware, explainable and multimodal fake news detection systems.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to HNB Garhwal University, Srinagar, Garhwal, Uttarakhand, India, for providing the necessary resources and institutional support for this research. Special gratitude is extended to the supervisor, Dr. Pritam Singh Negi, for his invaluable guidance, constructive support and mentorship throughout the research. His expertise and academic guidance were instrumental in shaping and completing this work.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, "Fake news detection on social media," *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, pp. 22-36, 2017.

[2] S. Vosoughi, D. Roy and S. Aral, "The spread of true and false news online," *Science*, Vol. 359, No. 6380, pp. 1146-1151, 2018.

[3] X. Zhou and R. Zafarani, "A survey of fake news," *ACM Comput. Surveys*, Vol. 53, No. 5, 2020. [Online]. Available: 10.1145/3395046

[4] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Inf. Sci.*, Vol. 497, pp. 38-55, 2019.

[5] L. Breiman, "Random forests," *Mach. Learn.*, Vol. 45, pp. 5-32, 2001.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, Vol. 20, pp. 273-297, 1995.

[7] D.-D. Konnon, E. A. Codjo, E. D. Togbe, G. Ogoudedji and R. Kounouewa, *Programme national de développement de la filière palmier a huile AU Bénin*, 2020.

[8] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, Vol. 1, pp. 81-106, 1986.

[9] J. C. S. Reis, A. Correia, F. Murai, A. Veloso and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intell. Syst.*, Vol. 34, No. 2, pp. 76-81, 2019.

[10] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti and M. Akbar, "Fake news detection using deep learning models: A novel approach," *Trans. Emerging Telecommun. Technol.*, Vol. 31, No. 2, 2019. [Online]. Available: 10.1002/ett.3767

[11] R. K. Kaliyar, A. Goswami and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools Appl.*, Vol. 80, pp. 11765-11788, 2021.

[12] R. K. Kaliyar, A. Goswami, P. Narang and S. Sinha, "FNDNet – A deep convolutional neural network for fake news detection," *Cognit. Syst. Res.*, Vol. 61, pp. 32-44, 2020.

[13] J. Alghamdi, S. Luo and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimedia Tools Appl.*, Vol. 83, pp. 51009-51067, 2023.

[14] C. Gunavathi and K. Premalatha, "Cuckoo search optimisation for feature selection in cancer classification: A new approach," *Int. J. Data Mining Bioinf.*, Vol. 13, No. 3, pp. 248-265, 2015.

[15] M. Zhao and Y. Qin, "Feature selection on elite hybrid binary cuckoo search in binary label classification," *Comput. Math. Methods Med.*, Vol. 2021, 2021. [Online]. Available: 10.1155/2021/5588385

[16] R. İncir, M. Yağanoğlu and F. Bozkurt, "A study on effective feature extraction and genetic algorithm based feature selection method in fake news detection classification using machine learning approaches," *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, Vol. 14, No. 3, pp. 764-776, 2024.

[17] F. Farhangian, R. M. O. Cruz and G. D. C. Cavalcanti, "Fake news detection: Taxonomy and comparative study," *Inf. Fusion*, Vol. 103, 2024. [Online]. Available: 10.1016/j.inffus.2023.102140

[18] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, Vol. 24, No. 5, pp. 513-523, 1988.

[19] C. Comito, L. Caroprese and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," *Social Network Anal. Min.*, Vol. 13, 2023. [Online]. Available: 10.1007/s13278-023-01104-w

[20] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata and R. Procter, "Detection and resolution of rumours in social media," *ACM Comput. Surv.*, Vol. 51, No. 2, 2018. [Online]. Available: 10.1145/3161603

[21] Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 19-23 Aug. 2018. New York, NY, USA: ACM, 2018, pp. 849-857.

[22] D. Khattar, J. S. Goud, M. Gupta and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW '19: The World Wide Web Conference*, 13-17 May. 2019. New York, NY, USA: ACM, 2019, pp. 2915-2921.

[23] K. Shu, S. Wang and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *WSDM '19: Proc. Twelfth ACM International Conference on Web Search and Data Mining*, 11-15 Feb. 2019. New York, NY, USA: ACM, 2019, pp. 312-320.

[24] N. Ruchansky, S. Seo and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *CIKM '17: Proc. 2017 ACM on Conference on Information and Knowledge Management*, 6-10 Nov. 2017. New York, NY, USA: ACM, 2017, pp. 797-806.

[25] F. Alam *et al.*, "A survey on multimodal Disinformation Detection," in *Proc. 29th International Conference on Computational Linguistics*, 12-17 Oct. 2022. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 6625-6643.

[26] K. Shu, L. Cui, S. Wang, D. Lee and H. Liu, "dEFEND: Explainable fake news detection," in *KDD '19: Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2019. New York, NY, USA: ACM, 2019, pp. 395-405.

[27] W. Y. Wang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 30 Jul. – 4 Aug. 2017. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 422-426.

[28] P. Patwa *et al.*, "Fighting an infodemic: Covid-19 fake news dataset," in *Proc. First International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT 2021)*, Feb. 2021. Springer International Publishing: Springer International Publishing, 2021, pp. 21-29.