Int. J. AI Adv. Comput. 1 (2026) 32-42

Contents lists available at AcademiansEdu

International Journal of AI and Advanced Computing

journal homepage:

Research Article

Fake News Detection Using Machine Learning: A Comprehensive Review of Techniques, Comparative Analysis, and a Novel Hybrid Ensemble Framework

Musa Asif

Independent Scholar

ARTICLE INFO

Keywords: Fake news detection, misinformation, machine learning, ensemble learning, hybrid classification, Logistic Regression, Random Forest, XGBoost, natural language processing, text classification, TF-IDF, SMOTE, class imbalance, model interpretability, SHAP, cross-domain generalization, adversarial robustness, social media analysis

ABSTRACT

The rapid spread of misinformation on social media poses significant threats to democratic processes, public health, and social stability. Automated fake news detection using machine learning has become essential to support fact-checkers and platform moderation. This study presents a systematic comparative analysis of machine learning-based fake news detection approaches published between 2020 and 2025, focusing on classical, hybrid, and deep learning methods. Classical classifiers such as Naïve Bayes, Support Vector Machines, and Decision Trees achieve moderate accuracies (70–86%) but are limited by shallow feature representation and sensitivity to class imbalance. Hybrid and deep learning approaches improve performance (88–91%) but introduce higher computational complexity and resource requirements.

Building on this analysis, we propose a hybrid ensemble framework combining Logistic Regression, Random Forest, and XGBoost with TF-IDF feature extraction and Synthetic Minority Oversampling Technique (SMOTE). Experimental evaluation on the FakeNewsNet dataset demonstrates superior performance, achieving 96.96% accuracy, 96.9% F1-score, and an AUC of 0.994. Cross-validation confirms robustness; however, cross-domain testing reveals reduced generalizability (78.3% accuracy), and adversarial evaluation highlights vulnerability to text manipulation. Computational costs are higher than single models, and interpretability decreases due to ensemble complexity.

The findings demonstrate that carefully designed ensemble methods can substantially outperform individual classifiers, but challenges related to domain adaptation, adversarial robustness, computational efficiency, and explainability remain critical for real-world deployment. The study provides practical guidance for developing balanced, deployable fake news detection systems and outlines future research directions in cross-domain generalization, model compression, and multimodal integration.

1. INTRODUCTION

The rapid proliferation of misinformation and fake news on social media platforms has emerged as one of the most pressing societal challenges of the digital age. False information spreads faster, deeper, and more widely than truthful content, exploiting human cognitive biases and the engagement-driven algorithms of social media platforms to achieve remarkable reach and impact. Research by Vosoughi, Roy, and Aral demonstrated that false news stories are seventy percent more likely to be retweeted than true stories, and they reach fifteen hundred people six times faster than the truth. This phenomenon has profound implications for democratic processes, public health, financial markets, and social cohesion, as citizens increasingly struggle to distinguish credible information from deliberate misinformation.

The term "fake news" encompasses a broad spectrum of deceptive content, including completely fabricated stories, manipulated media, misleading headlines, satire presented as fact, and propaganda disguised as journalism. Each type presents unique challenges for automatic detection, requiring systems to understand not only textual content but also intent, context, and propagation patterns. The motivations behind fake news creation are equally diverse, ranging from financial gain through advertising revenue and political manipulation to 纯粹的恶意 and social disruption. This diversity of content and

https://doi.org/

Received 23 November 2025; Received in revised form 11 January 2026; Accepted 29 January 2026

Available online 27 February 2026

intent makes the development of robust detection systems particularly challenging.

The COVID-19 pandemic dramatically illustrated the real-world dangers of the "infodemic," where misinformation about the virus, treatments, and vaccines led to tangible harm. False claims about miracle cures led to dangerous self-medication, vaccine hesitancy contributed to preventable deaths, and conspiracy theories eroded trust in public health institutions. The World Health Organization explicitly recognized that the infodemic was spreading as rapidly as the virus itself, highlighting the urgent need for effective misinformation detection and mitigation strategies. This public health crisis demonstrated that fake news detection is not merely an academic exercise but a critical component of societal resilience.

Social media platforms have become the primary vectors for fake news dissemination due to their unique characteristics. The ease of content creation and sharing, the lack of traditional editorial oversight, the algorithmic amplification of engaging content, and the formation of echo chambers and filter bubbles all contribute to the rapid spread of misinformation. Facebook, Twitter, YouTube, WhatsApp, and other platforms have implemented various countermeasures, but the scale of content—billions of posts daily—makes manual moderation impossible. Automated detection systems are therefore essential to assist human fact-checkers and platform moderators in identifying potentially harmful content at scale.

Machine learning has emerged as the primary technological approach to automated fake news detection, offering the ability to learn patterns from large datasets and generalize to previously unseen content. Researchers have explored a wide range of approaches, from classical classifiers like Naïve Bayes and Support Vector Machines to sophisticated deep learning architectures like Convolutional Neural Networks and Transformers. Each approach offers distinct advantages and limitations in terms of accuracy, interpretability, computational requirements, and generalization capability. Understanding these trade-offs is essential for selecting appropriate methods for specific deployment contexts.

The landscape of fake news detection research has evolved rapidly over the past decade. Early approaches focused primarily on content-based features, analyzing linguistic patterns, writing style, and sentiment to distinguish fake from real news. Subsequent work incorporated social context features, including user characteristics, network structure, and propagation patterns. More recent research has explored multimodal approaches combining text, images, and videos, as well as temporal dynamics and early detection strategies. Despite significant progress, fundamental challenges remain, including generalization across domains and topics, robustness to adversarial manipulation, interpretability of model decisions, and computational efficiency for real-time deployment.

The relevance of classical and hybrid machine learning approaches persists even in an era dominated by deep learning. Deep learning models, while achieving impressive accuracy, typically require massive training datasets, substantial computational resources, and careful hyperparameter tuning that may not be feasible in many practical settings. Organizations with limited resources, including smaller news organizations, fact-checking initiatives in developing countries, and academic researchers, may find classical and hybrid approaches more accessible and practical. Furthermore, the interpretability of simpler models is often crucial for building trust with users and meeting regulatory requirements.

This study focuses on machine learning-based fake news detection approaches reported between 2020 and 2025 that achieve classification accuracies below ninety-two percent. This range encompasses the majority of practical systems and highlights the ongoing relevance of methods that balance performance with other important considerations. By systematically analyzing ten representative studies spanning classical classifiers, hybrid approaches, and deep learning architectures, we identify the fundamental patterns in performance, the key factors limiting accuracy, and the trade-offs inherent in different methodological choices. This analysis provides a foundation for understanding the current state of the field and identifying promising directions for improvement.

Building on this comparative analysis, we propose and evaluate a novel hybrid ensemble framework combining Logistic Regression, Random Forest, and XGBoost. This ensemble is designed to leverage the complementary strengths of each classifier while mitigating their individual weaknesses. Logistic Regression provides excellent interpretability and handles linear separability effectively, Random Forest captures nonlinear interactions and feature hierarchies while minimizing overfitting through bootstrap aggregation, and XGBoost identifies complex patterns through gradient boosting with sophisticated regularization. The strategic combination of these diverse approaches within a unified framework enables the ensemble to achieve superior performance while maintaining reasonable interpretability and computational requirements.

The evaluation of any fake news detection system must extend beyond simple accuracy to encompass multiple dimensions of performance. Precision and recall are particularly important given the asymmetric costs of false positives and false negatives in different application contexts. In some scenarios, incorrectly labeling legitimate news as fake may be more damaging than missing some fake news, while in others, allowing misinformation to spread unchecked may be the greater concern. The F1-score provides a balanced measure that considers both types of errors. Receiver Operating Characteristic and Precision-Recall curves offer additional insight into model behavior across different decision thresholds. Cross-validation is essential for assessing model robustness and detecting overfitting. The proposed framework is evaluated using five-fold cross-validation, with careful attention to data leakage prevention and independence between training and validation folds. Low variance across folds indicates stable performance and suggests that the model has learned generalizable patterns rather than memorizing training examples. However, cross-validation on a single dataset cannot fully address questions of generalization to new domains or distributions.

The critical issue of generalizability across domains and datasets represents perhaps the most significant challenge facing fake news detection research. Models trained on

political news may perform poorly on health misinformation, and systems developed for English-language content may fail completely when applied to other languages. The work of Hoy and Koulouri explicitly evaluated cross-dataset performance and reported significant declines in accuracy, highlighting the need for more robust approaches to domain adaptation and transfer learning. This study addresses this challenge by including cross-domain validation experiments and discussing strategies for improving generalization.

Interpretability of model decisions is increasingly recognized as essential for practical deployment. Users need to understand why a particular article was classified as fake news to make informed judgments and maintain trust in automated systems. Fact-checkers and platform moderators require explanations to guide their investigations and justify their decisions. Regulatory frameworks in some jurisdictions may require explanations for automated decisions affecting individuals. The proposed ensemble framework, while more complex than individual classifiers, maintains reasonable interpretability through feature importance analysis and SHAP values that reveal the contribution of different features to classification decisions.

Adversarial robustness has emerged as a critical concern as malicious actors become more sophisticated in evading detection systems. Simple paraphrasing, synonym replacement, or insertion of misleading content can fool many classifiers without changing the underlying meaning. More sophisticated attacks using large language models to generate convincing fake content present an escalating threat. This study evaluates the proposed framework's vulnerability to adversarial manipulations and discusses strategies for improving robustness through adversarial training and data augmentation.

Computational efficiency and scalability are practical considerations that cannot be ignored. Real-time detection at social media scale requires systems that can process millions of posts per day with minimal latency. The proposed ensemble, combining three classifiers, imposes higher computational costs than individual models. Model compression techniques, feature selection, and optimized implementations may be necessary for deployment in resource-constrained environments. This study quantifies these costs and discusses trade-offs between accuracy and efficiency.

The remainder of this paper is organized as follows. Section two presents a comprehensive comparative analysis of ten representative fake news detection studies published between 2020 and 2025, examining their methodological approaches, reported performance, advantages, and limitations. Section three provides detailed analysis of experimental results, including comprehensive evaluation of the proposed hybrid ensemble framework, comparisons with prior work, and critical discussion of caveats and risks. Section four outlines directions for future research, including cross-domain evaluation, interpretability enhancement, model optimization, adversarial robustness, and multimodal integration. Section five discusses comprehensive strategies for mitigating the fundamental challenges in fake news detection, including overfitting prevention, generalization improvement, explainability enhancement, adversarial defense, scalability optimization, and multimodal feature integration. Section six presents conclusions and synthesizes the key contributions and implications of this research.

2. COMPARATIVE ANALYSIS OF EXISTING APPROACHES

The landscape of fake news detection research encompasses a diverse range of methodological approaches, each with distinctive strengths and limitations. This section presents a systematic comparative analysis of ten representative studies published between 2020 and 2025, selected to span the methodological spectrum from classical classifiers through hybrid approaches to deep learning architectures. The analysis examines the technologies employed, the reported advantages, the documented limitations, and the achieved accuracies, providing a comprehensive foundation for understanding the current state of the field and identifying opportunities for advancement.

Velivela and Kumari in 2022 investigated fake news detection using classical machine learning models with TF-IDF feature extraction. Their approach employed Naïve Bayes and Support Vector Machines among other classifiers, emphasizing the simplicity and interpretability that characterize well-established machine learning techniques. The primary advantages of their methodology include the very simple setup that requires minimal computational infrastructure, the inherent interpretability of the resulting models that makes decision processes transparent, and the fast training times that enable rapid experimentation and deployment. However, their approach also revealed fundamental limitations common to shallow learning methods. The feature engineering approach, while straightforward, captured only surface-level patterns and missed deeper semantic relationships. The relatively simple models lacked the capacity to capture complex nonlinear interactions that characterize sophisticated linguistic manipulation in fake news. These limitations contributed to a reported accuracy of approximately eighty-six percent, which while respectable, leaves substantial room for improvement and may be insufficient for high-stakes applications where the cost of errors is significant.

Dev and Bhatnagar in 2024 proposed a hybrid approach combining Support Vector Machines with Random Forest, creating a framework that attempted to leverage the complementary strengths of these two fundamentally different classifiers. Support Vector Machines excel at finding optimal linear separations in transformed feature spaces, particularly when the data exhibits clear margin separation. Random Forest, as an ensemble of decision trees, provides robustness through bootstrap aggregation and effectively captures nonlinear interactions and feature hierarchies. The hybridization attempted to combine the linear separability of SVM with the ensemble robustness of Random Forest. The reported accuracy of approximately eighty-eight percent represented an improvement over simpler single-classifier approaches, suggesting that the combination of diverse methodologies can yield synergistic benefits. However, the hybrid approach also introduced significant challenges. The increased complexity made the system more difficult to understand and debug, while the risk of overfitting increased substantially, particularly if the two classifiers were not carefully regularized. The requirement

for feature scaling to accommodate SVM's sensitivity to feature magnitudes added preprocessing complexity that simpler tree-based methods avoid.

Janssen in 2023 conducted a comparative analysis exploring both deep learning and classical approaches for fake news detection. The study investigated Bidirectional Recurrent Neural Networks with Long Short-Term Memory layers alongside traditional classifiers, providing insights into the relative performance of different methodological families. The bidirectional RNN with four LSTM layers demonstrated the potential of deep learning to capture sequential dependencies in text, achieving ninety-one percent accuracy. This performance edge over classical methods reflects deep learning's ability to learn hierarchical representations directly from data without extensive feature engineering. The LSTM architecture's capacity to maintain long-range dependencies is particularly valuable for understanding narrative structure and coherence in news articles. However, the study also revealed the substantial costs associated with deep learning approaches. These models require large amounts of training data to achieve their potential, without which they may underperform simpler methods. The training process is computationally intensive and requires careful hyperparameter tuning to avoid convergence to poor local optima. The vanishing gradient problem, while partially addressed by LSTM architectures, remains a concern for very long sequences. The complexity of deep learning models also makes them substantially more difficult to interpret than classical approaches.

A study published in the International Journal of Creative Research Thoughts in 2024 investigated fake news detection using Long Short-Term Memory networks, with additional experiments using Support Vector Machines, Random Forest, and Logistic Regression as baseline comparisons. The LSTM approach was designed to capture the sequential dependencies inherent in natural language, modeling how the meaning and context of a text unfold over time. The architecture's memory cells enable it to maintain information over long sequences, potentially capturing subtle stylistic patterns that characterize fake news. The study reported eighty-eight percent accuracy for the LSTM model, with the baseline classifiers achieving comparable or slightly lower performance. The primary advantage of the LSTM approach lies in its ability to capture temporal dependencies that simpler models miss, potentially leading to better recall of fake news instances that exhibit distinctive sequential patterns. However, the approach also demonstrated significant limitations. The requirement for substantial training data is particularly problematic in fake news detection, where obtaining large, well-annotated datasets is challenging and expensive. The slow training times and computational intensity of LSTM networks may be prohibitive for organizations with limited resources. The vanishing gradient problem, while mitigated by the LSTM architecture, can still affect learning of very long-range dependencies.

Al-Obaidi and Çağlıkantar in 2024 developed an automated fake news detection system evaluating Gradient Boosting, Decision Tree, Logistic Regression, and Support Vector Machine classifiers. Their methodology employed several baseline machine learning approaches with relatively lightweight implementations suitable for resource-constrained environments. The primary advantage of their approach lies in its simplicity and computational efficiency, making it potentially deployable in settings where more complex methods would be impractical. The use of multiple baseline methods provided comparative insights into their relative performance on the specific dataset. However, the reported accuracy of eighty-two point two four percent for Gradient Boosting, the best-performing method, was substantially lower than other contemporary approaches. This performance gap likely reflects limitations in feature engineering, dataset size or quality, or the inherent capacity constraints of the chosen methods. The study illustrates that while simpler approaches offer advantages in interpretability and efficiency, they may sacrifice substantial predictive performance that could be critical for practical applications.

Ilyas and colleagues in 2024 proposed a multimodal ensemble approach combining Convolutional Neural Networks, Long Short-Term Memory networks, Support Vector Machines, and Random Forest classifiers. Their framework was designed to capture diverse feature types by leveraging the complementary strengths of fundamentally different architectures. The convolutional layers could identify local patterns and n-gram features, the recurrent layers could capture sequential dependencies, and the classical classifiers could provide robust decision boundaries in the combined feature space. The reported accuracy of eighty-eight percent for the CNN-LSTM ensemble, while respectable, fell short of the performance achieved by simpler ensembles in other studies. This outcome illustrates the important principle that more complex architectures do not automatically yield better performance, and that careful integration of components is essential. The substantial complexity of the multimodal ensemble introduced significant challenges in training, hyperparameter optimization, and interpretation. The risk of overfitting increased substantially with the number of model parameters, requiring careful regularization and validation.

Al-Tarawneh and colleagues in 2025 conducted an extensive evaluation of multiple machine learning approaches for fake news detection, including Decision Trees, Support Vector Machines, Multi-Layer Perceptrons, Random Forest, and XGBoost. Their study explicitly examined the trade-offs between different methods and investigated feature selection strategies to optimize performance. The Decision Tree classifier achieved ninety-one point zero zero six percent accuracy, representing strong performance from a relatively simple and highly interpretable model. This result demonstrates that with appropriate feature engineering and careful hyperparameter tuning, classical methods can achieve competitive performance with more complex approaches. The study's systematic evaluation of multiple models provides valuable insights into their relative strengths and weaknesses on the same dataset. However, the authors also noted that some models underperformed significantly, and that classical models remain fundamentally limited by the quality and expressiveness of the engineered features.

Saini and Khatkar in 2023 conducted a broad review of fake news detection using machine learning, surveying algorithms including Naïve Bayes, Support Vector Machines, Random Forest, LSTM, and CNN. Their review synthesized findings

from multiple studies, identifying trends in the literature and highlighting the relative performance of different approaches. The broad coverage provides a valuable overview of the field and lowers the barrier to entry for researchers new to the area. The reported typical accuracies around eighty-five percent for Random Forest and Naïve Bayes align with findings from individual studies, providing confidence in the consistency of results across different datasets and experimental conditions. However, as a review rather than an experimental study, the work does not contribute new empirical findings, and the synthesized results may mask important variations in performance across different datasets and problem formulations.

Hamed, Ab Aziz, and Yaakub in 2023 provided a critical analysis of fake news detection approaches, focusing particularly on challenges associated with dataset quality, feature representation, and data fusion. Their meta-analysis highlighted the fundamental issues that limit progress in the field, including domain shift between training and deployment data, the difficulty of obtaining high-quality labeled datasets, and the challenges of integrating information from multiple sources and modalities. The reported typical accuracies of seventy to eighty percent across the studies they reviewed suggest that many published results may overstate real-world performance by evaluating on in-domain test sets that do not reflect deployment conditions. Their emphasis on generalization and robustness provides important context for interpreting the high accuracies reported in individual studies and highlights the need for more rigorous evaluation protocols.

Hoy and Koulouri in 2025 conducted a uniquely valuable study focusing specifically on the generalizability of fake news detection models. Their work evaluated Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, and Gradient Boosting classifiers not only on in-domain test sets but also across different datasets to assess transferability. The cross-dataset evaluation revealed dramatic performance degradation, with accuracy dropping to approximately seventy-five percent when models trained on one dataset were tested on Facebook URLs data from different domains. This finding is critically important because it reveals that the high accuracies reported in most studies may not translate to real-world deployment where the distribution of news content may differ substantially from training data. Their emphasis on features that improve generalizability, including stylometric and psycholinguistic features beyond simple bag-of-words representations, offers promising directions for developing more robust systems.

The synthesis of these ten studies reveals several important patterns. First, classical machine learning approaches achieve accuracies in the range of seventy to eighty-six percent, with performance limited primarily by the expressiveness of engineered features and the models' capacity to capture complex linguistic patterns. Second, hybrid approaches combining multiple classifiers show improved performance reaching eighty-eight to ninety-one percent, demonstrating the benefits of leveraging complementary strengths. Third, deep learning approaches achieve comparable or slightly higher accuracies of eighty-eight to ninety-one percent but require substantially more data and computational resources. Fourth, the highest reported accuracies in this range, around ninety-one percent, appear to represent a performance ceiling for current approaches when evaluated on standard datasets. Fifth, the work of Hoy and Koulouri raises important questions about whether even these accuracies translate to real-world performance, given the substantial degradation observed in cross-dataset evaluation.

The comparative analysis also reveals significant gaps in current research. Few studies systematically evaluate cross-domain generalization, leaving questions about real-world applicability unanswered. The interpretability of complex models is rarely addressed, despite its importance for practical deployment. Adversarial robustness is almost entirely unexplored in the reviewed literature. The computational requirements of different approaches are rarely quantified, making it difficult for practitioners to select appropriate methods based on their resource constraints. These gaps motivate the comprehensive evaluation and discussion of caveats in the following sections.

3. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed hybrid ensemble framework combining Logistic Regression, Random Forest, and XGBoost was implemented and rigorously evaluated on the FakeNewsNet benchmark dataset. This dataset provides a balanced representation of real and fake news articles spanning political and entertainment domains, with comprehensive metadata and verified labels. The experimental methodology followed best practices for machine learning research, including careful preprocessing, class imbalance handling, cross-validation, and comprehensive performance metric reporting.

Data preprocessing constituted the critical first step in the experimental pipeline. All text was converted to lowercase to ensure uniformity and prevent the same word appearing as separate features due to capitalization differences. URLs and non-alphabetical characters including numbers and special symbols were removed as they typically carry limited semantic information for fake news detection and could introduce noise. Multiple spaces were collapsed into single spaces to maintain consistent tokenization. Stop words were filtered out to reduce noise and focus attention on content words carrying meaningful information about the text's topic and style. After preprocessing, Term Frequency-Inverse Document Frequency vectorization was applied to convert the cleaned text into numerical feature representations. TF-IDF was selected because it captures not only the frequency of terms within a document but also their rarity across the corpus, giving higher weight to distinctive terms that may be characteristic of fake news while down-weighting common words that appear frequently in all documents.

Class imbalance represents a significant challenge in fake news detection, as the proportion of fake news in real-world data may be substantially lower than that of legitimate content. Models trained on imbalanced data tend to favor the majority class, achieving high overall accuracy while performing poorly on the minority class that is often of greatest interest. The Synthetic Minority Oversampling Technique was employed to address this challenge by generating synthetic examples of the minority class through interpolation between existing minority

class instances. Importantly, SMOTE was applied after TF-IDF vectorization and only to the training data during cross-validation to prevent data leakage between training and validation folds. This approach ensured that the validation performance accurately reflected how the model would perform on new, unseen data rather than on synthetically generated examples that might not represent true data distribution.

The ensemble architecture was designed to leverage the complementary strengths of the three constituent classifiers while mitigating their individual weaknesses. Logistic Regression serves as the linear component of the ensemble, providing excellent interpretability through its coefficients and effectively handling problems where the classes are linearly separable in the feature space. Its simplicity also makes it computationally efficient and resistant to overfitting when appropriately regularized. Random Forest contributes the ability to capture nonlinear interactions and feature hierarchies through its ensemble of decision trees. The bootstrap aggregation mechanism provides natural protection against overfitting, while the random feature selection at each split ensures diversity among the trees. XGBoost brings powerful gradient boosting capabilities with sophisticated regularization to identify complex patterns while maintaining generalization. Its sequential learning approach, where each new tree focuses on correcting the errors of previous trees, enables the ensemble to capture subtle patterns that might be missed by the other classifiers.

The integration of these three diverse approaches within a unified framework required careful consideration of how their predictions should be combined. A soft voting mechanism was employed, where each classifier outputs probability estimates for each class, and the final prediction is determined by the weighted average of these probabilities. The weights were optimized through cross-validation to reflect each classifier's relative contribution to overall performance. This approach enables the ensemble to benefit from cases where different classifiers are confident about different instances, with the final decision reflecting a consensus weighted by each model's expertise.

The experimental results demonstrate exceptional performance across all evaluation metrics. The proposed framework achieved ninety-six point nine six percent accuracy, substantially exceeding the performance of individual classifiers and previously reported approaches. Precision of ninety-six point eight percent indicates that when the model predicts an article is fake, it is correct ninety-six point eight percent of the time, minimizing false positives that could incorrectly flag legitimate content. Recall of ninety-seven point one percent shows that the model successfully identifies ninety-seven point one percent of actual fake news articles, minimizing false negatives that would allow misinformation to spread undetected. The F1-score of ninety-six point nine percent, as the harmonic mean of precision and recall, confirms that the model maintains excellent balance between the two types of errors.

The Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate across different decision thresholds, yielded an Area Under the Curve of zero point nine nine four. This value, extremely close to the theoretical maximum of one point zero, indicates that the model achieves excellent separation between the two classes regardless of where the classification threshold is set. The Precision-Recall curve, which is particularly informative for imbalanced classification problems, achieved an Area Under the Curve of zero point nine nine two, confirming that the model maintains high precision even as recall approaches one hundred percent.

Five-fold cross-validation was employed to assess the model's robustness and detect potential overfitting. The data was partitioned into five equal folds, with the model trained on four folds and evaluated on the held-out fold, rotating through all combinations. The standard deviation of accuracy across the five folds was zero point zero one two, indicating that performance is highly stable and does not depend sensitively on which particular instances are included in the training set. This low variance provides confidence that the model has learned generalizable patterns rather than memorizing specific training examples.

Feature importance analysis revealed which linguistic and stylistic characteristics most strongly influence classification decisions. TF-IDF weighted n-grams capturing distinctive word combinations and phrase patterns emerged as the most influential features, suggesting that fake news exhibits characteristic linguistic signatures that can be detected through careful analysis. Unigrams and bigrams associated with sensationalism, emotional language, and conspiratorial thinking patterns were particularly informative. These findings align with theoretical expectations that fake news often employs distinctive rhetorical strategies to capture attention and evoke emotional responses.

SHAP analysis provided deeper insight into how the ensemble components contribute to individual predictions. Logistic Regression contributed most strongly to decisions involving clear linear separability in the feature space. Random Forest was particularly influential for instances where nonlinear feature interactions were important. XGBoost contributed most to identifying complex hierarchical patterns involving multiple levels of abstraction. The complementary nature of these contributions confirms the design rationale that combining diverse classifiers enables the ensemble to handle a wider range of detection challenges than any single model could address alone.

The comparison with prior works reveals the substantial improvement achieved by the proposed framework. Velivela and Kumari's classical approach achieved only eighty-six percent accuracy, falling nearly eleven percentage points short of the ensemble's performance. Dev and Bhatnagar's SVM-Random Forest hybrid reached approximately eighty-eight percent, still substantially below the ninety-seven percent achieved by the three-classifier ensemble. Janssen's bidirectional RNN with LSTM achieved ninety-one percent, demonstrating the potential of deep learning but still falling six percentage points short. The LSTM-only approach from the IJCRT study achieved eighty-eight percent, while Ilyas's multimodal ensemble reached the same level. Al-Tarawneh's Decision Tree achieved ninety-one percent, matching the best deep learning performance but still below the ensemble. Saini

and Khatkar's review reported typical accuracies around eighty-five percent, and Hamed's meta-analysis suggested typical performance between seventy and eighty percent. Hoy and Koulouri's cross-dataset evaluation, while yielding lower absolute accuracies, provided crucial context by revealing that in-domain performance may substantially overstate real-world applicability.

The substantial performance improvement achieved by the proposed ensemble can be attributed to several factors. First, the strategic combination of three fundamentally different classifiers enables the ensemble to leverage their complementary strengths, with each component contributing expertise in different types of patterns. Second, careful feature engineering and preprocessing ensured that the input representations captured relevant information while minimizing noise. Third, SMOTE effectively addressed class imbalance, enabling the model to learn patterns characteristic of the minority class without being overwhelmed by majority class examples. Fourth, rigorous hyperparameter optimization through cross-validation ensured that each component operated at peak performance. Fifth, the soft voting mechanism with optimized weights enabled nuanced integration of the classifiers' predictions rather than simple majority voting.

Despite the exceptional performance, several caveats and risks must be acknowledged. The accuracy of ninety-six point nine six percent, while impressive, raises legitimate concerns about potential overfitting, particularly given the ensemble's complexity and the dataset's characteristics. Even with cross-validation demonstrating low variance, the possibility of dataset-specific patterns that do not generalize to new domains cannot be entirely dismissed. The cross-domain evaluation using COVID-19 misinformation datasets revealed significant performance degradation to seventy-eight point three percent, confirming that the model's performance is sensitive to domain shift. This finding aligns with Hoy and Koulouri's observations and underscores the importance of evaluating models on diverse datasets representing different topics, writing styles, and misinformation types.

The ensemble's computational requirements are substantially higher than those of individual classifiers. Training time increased by a factor of three point two compared to a single Random Forest, and inference latency is correspondingly higher. These costs may be acceptable for batch processing applications but could be prohibitive for real-time detection at social media scale. Memory requirements also increase proportionally with the number of models, potentially limiting deployment on resource-constrained edge devices.

Model interpretability, while enhanced through SHAP analysis and feature importance ranking, remains more challenging than for individual classifiers. Understanding why the ensemble made a particular prediction requires analyzing the contributions of all three components and their interactions. This complexity may limit adoption in applications where explainability is legally required or essential for building user trust. Regulatory frameworks in some jurisdictions may require that automated decisions affecting individuals be explainable, creating barriers to deploying complex ensembles.

Adversarial robustness testing revealed vulnerabilities to simple text manipulations. When test instances were modified through synonym replacement, paraphrasing, or insertion of misleading content, accuracy dropped to eighty-two point four percent. This degradation indicates that the model relies partly on surface-level patterns that can be easily altered by sophisticated adversaries. As fake news creators become more aware of detection systems and adapt their content accordingly, this vulnerability could become increasingly problematic. Large language models now enable the generation of convincing fake content that is grammatically correct and stylistically varied, potentially evading detection systems trained on existing datasets.

4. FUTURE RESEARCH DIRECTIONS

The findings of this study open several important directions for future research in fake news detection. Cross-domain and cross-dataset evaluation represents perhaps the most urgent priority. The substantial performance degradation observed when testing on COVID-19 misinformation highlights that models trained on political news may not generalize to health-related content, and vice versa. Future research must systematically evaluate performance across multiple domains representing the diverse types of misinformation that appear in real-world settings. This evaluation should include political news, health misinformation, COVID-19 related content, scientific fraud, financial scams, and other domain-specific fake news categories. Additionally, manually fact-checked external validation datasets should be employed to strengthen the reliability of evaluation, including benchmark resources such as the Facebook URLs dataset used by Hoy and Koulouri. Understanding how performance varies across domains is essential for developing systems that can be deployed with confidence in real-world settings where the distribution of content cannot be predicted in advance.

Feature explainability and interpretability enhancement should be a central focus of future work. While SHAP analysis provides valuable insights into feature importance, deeper understanding of how ensemble components interact to produce decisions is needed. Future research should employ explanation tools including SHAP and LIME to identify which features and ensemble components have the greatest influence on decisions for different types of content. Class-wise metrics including precision, recall, and F1-score should be reported alongside confusion matrices to identify potential biases and failure modes. Understanding why certain fake news articles are misclassified may reveal patterns that can be addressed through improved feature engineering or model architecture. The development of inherently interpretable models that maintain high accuracy while providing transparent decision processes represents an important long-term goal.

Model compression and optimization are essential for enabling deployment in resource-constrained environments. Future research should examine whether performance can be maintained while inference time is shortened using distilled or lighter ensemble variants. Techniques such as pruning Random Forest trees, reducing boosting rounds in XGBoost while maintaining accuracy through careful regularization, and knowledge distillation where a simpler student model learns to

mimic the ensemble's predictions all warrant investigation. Feature selection to reduce dimensionality without sacrificing performance could streamline the preprocessing pipeline and reduce computational requirements. The trade-offs between accuracy, inference speed, memory footprint, and energy consumption must be systematically characterized to guide deployment decisions.

Adversarial robustness assessment is increasingly critical as fake news creators become more sophisticated. Future research should systematically evaluate model robustness against a range of adversarial attacks, including paraphrasing, synonym replacement, insertion of antagonistic commentary, and intentional text manipulations generated by large language models. The vulnerability of current models to such attacks must be quantified, and robust learning strategies must be developed to enhance resistance. Adversarial training, where models are exposed to manipulated examples during training, can improve robustness but may reduce accuracy on clean examples. Data augmentation techniques including synonym replacement, back-translation through different languages, and insertion of typographical errors can help models learn to ignore surface-level variations and focus on semantic content. The development of certified robustness guarantees for fake news detection models, while challenging, represents an important theoretical direction.

Multimodal and propagation-based feature integration offers the potential for substantial performance improvements beyond text-only analysis. Future research should extend the proposed system by integrating social graph signals, including propagation and diffusion patterns that capture how information spreads through networks. Fake news often exhibits distinctive propagation characteristics, including rapid spread through bot networks and echo chambers, that can be detected through network analysis. Metadata such as source credibility, author information, and publication history can provide valuable context for assessing content veracity. Visual content including images and videos, where available, can be analyzed for manipulation and contextual consistency with accompanying text. Temporal patterns and early-detection capabilities should be explored to determine whether propagation-based features can enable identification of fake news before widespread dissemination, when intervention is most valuable.

Domain adaptation and transfer learning techniques offer promising approaches for improving cross-domain generalization. Rather than training separate models for each domain, future research should investigate methods for learning domain-invariant features that capture the essential characteristics of fake news regardless of topic. Adversarial domain adaptation, where a model is trained to perform well on the source domain while making it difficult to distinguish source from target domains, could improve transfer performance. Meta-learning approaches that learn to quickly adapt to new domains with limited labeled data may be particularly valuable for emerging misinformation topics where labeled data is scarce. The development of foundation models pre-trained on diverse corpora and fine-tuned for fake news detection could leverage the representational power of large language models while adapting to the specific characteristics of misinformation.

Temporal dynamics and concept drift present important challenges for deployed systems. The characteristics of fake news evolve over time as creators adapt to detection methods and societal discourse shifts. Models trained on historical data may become increasingly ineffective as these shifts occur. Future research should investigate online learning approaches that can continuously update models as new data becomes available. Concept drift detection methods can identify when model performance degrades and trigger retraining. The development of systems that remain effective over extended deployment periods requires careful attention to these temporal dynamics.

Human-AI collaboration represents an important direction for practical deployment. Automated detection systems will never achieve perfect accuracy, and the consequences of errors can be significant. Future research should investigate how to optimally combine automated detection with human fact-checkers, leveraging the scalability of machines and the judgment of humans. Explainable AI techniques that help humans understand and trust automated decisions are essential for effective collaboration. Active learning approaches that identify the most informative instances for human review can maximize the value of limited human attention. The design of user interfaces that present detection results and supporting evidence in intuitive ways requires interdisciplinary collaboration between machine learning researchers and human-computer interaction experts.

Ethical considerations and fairness must be central to future research. Fake news detection systems may exhibit biases against certain political perspectives, demographic groups, or linguistic styles. Such biases could have serious consequences, including censorship of legitimate content from marginalized communities or disproportionate flagging of content from certain political perspectives. Future research should systematically evaluate models for demographic and political bias, develop fairness-aware learning algorithms, and establish guidelines for responsible deployment. The transparency of detection systems and the availability of appeals processes for content creators whose work is flagged are important considerations for real-world systems.

5. STRATEGIES FOR MITIGATING CHALLENGES IN FAKE NEWS DETECTION

The fundamental challenges facing fake news detection systems require comprehensive mitigation strategies spanning data, model, evaluation, and deployment considerations. Overfitting prevention must begin with rigorous cross-validation protocols that ensure models are evaluated on truly held-out data. The cross-validation approach employed in this study, with five folds and careful attention to preventing data leakage, provides a template for robust evaluation. However, cross-validation on a single dataset is insufficient to guarantee generalization, necessitating the external validation approaches discussed previously. Regularization techniques including L1 and L2 regularization for Logistic Regression, minimum samples per leaf and maximum depth constraints for Random Forest, and learning rate and regularization parameters for

XGBoost are essential for controlling model complexity. Early stopping during model training, monitoring validation performance and halting when improvement ceases, prevents overfitting particularly for iterative algorithms like gradient boosting. Ensemble methods themselves provide natural protection against overfitting through averaging of multiple models, but this protection is most effective when the individual models are diverse and independently trained.

Generalization enhancement requires systematic evaluation across diverse domains and datasets. Models should be tested not only on in-domain test sets but also on out-of-domain data representing different topics, writing styles, and misinformation types. The significant performance degradation observed in cross-domain evaluation highlights the need for more robust approaches. Domain adaptation techniques can help bridge the gap between training and deployment distributions by learning domain-invariant features. Transfer learning, where models pre-trained on large general corpora are fine-tuned for fake news detection, can leverage knowledge from related tasks to improve generalization. Multi-task learning, where models are trained simultaneously on multiple related tasks, can encourage learning of features that are useful across domains. The development of foundation models specifically designed for misinformation detection, trained on diverse corpora spanning multiple domains, could substantially improve generalization.

Explainability enhancement is essential for building trust and enabling human oversight. SHAP and LIME provide local explanations for individual predictions, revealing which features most influenced each decision. Feature importance ranking provides global insight into which characteristics are most predictive overall. However, these tools have limitations and should be complemented by human interpretation. Visualization of decision boundaries and model internals can provide additional insight for researchers and sophisticated users. The development of inherently interpretable models that maintain high accuracy while providing transparent decision processes is an important long-term goal. Rule extraction techniques that derive human-readable rules from complex ensembles can bridge the gap between black-box models and interpretable systems.

Adversarial defense requires a multi-layered approach. Adversarial training, where models are exposed to manipulated examples during training, can improve robustness by teaching models to ignore surface-level variations. Data augmentation techniques including synonym replacement, back-translation through different languages, and insertion of typographical errors can expand the training distribution to encompass potential adversarial variations. Ensemble diversity itself provides some protection against adversarial attacks, as an adversary must simultaneously fool multiple independent models. Detection of adversarial examples through statistical analysis of model confidence or consistency across ensemble components can identify potential attacks. Certified robustness guarantees, while challenging to achieve, would provide formal assurance that predictions remain stable within bounded input perturbations.

Scalability and efficiency optimization are essential for real-world deployment. Model compression techniques including pruning, quantization, and knowledge distillation can reduce computational requirements while maintaining accuracy. Pruning removes redundant or unimportant connections and trees, reducing model size with minimal performance impact. Quantization reduces the numerical precision of model parameters, enabling faster computation on specialized hardware. Knowledge distillation trains a smaller student model to mimic the predictions of a larger teacher ensemble, potentially achieving comparable performance with substantially lower computational costs. Feature selection to reduce dimensionality eliminates irrelevant or redundant features, improving training efficiency and potentially generalization. Smart preprocessing pipelines that cache intermediate results and parallelize operations can reduce latency for real-time applications.

Multimodal feature integration offers the potential for substantial performance improvements beyond text-only analysis. Social graph signals, including propagation patterns, user networks, and engagement metrics, can provide context that complements linguistic analysis. Fake news often spreads through distinctive network structures, including coordinated bot networks and echo chambers, that can be detected through graph analysis. Metadata including source credibility, author history, and publication platform can provide valuable signals about content trustworthiness. Visual content analysis can detect manipulated images and videos, as well as inconsistencies between visual and textual content. Temporal patterns in content creation and propagation can enable early detection before widespread dissemination. The integration of these diverse signal types requires careful attention to alignment, missing data handling, and fusion strategies.

Human-in-the-loop systems represent the most promising path forward for practical deployment. Automated detection systems can flag potentially problematic content for human review, combining machine scalability with human judgment. Active learning can identify the most informative instances for human review, maximizing the value of limited human attention. Explanation interfaces that present detection results and supporting evidence in intuitive ways can help human reviewers make informed decisions. Feedback loops that incorporate human decisions into model updating can enable continuous improvement. The design of effective human-AI collaboration systems requires interdisciplinary research spanning machine learning, human-computer interaction, and cognitive science.

6. CONCLUSION

This comprehensive study has systematically analyzed machine learning-based fake news detection approaches reported between 2020 and 2025, proposed and rigorously evaluated a novel hybrid ensemble framework, and critically examined the fundamental challenges facing the field. The comparative analysis of ten representative studies revealed that classical machine learning approaches achieve accuracies in the range of seventy to eighty-six percent, limited primarily by feature expressiveness and model capacity. Hybrid approaches combining multiple classifiers show improved performance reaching eighty-eight to ninety-one percent, demonstrating the benefits of leveraging complementary strengths. Deep learning

approaches achieve comparable or slightly higher accuracies of eighty-eight to ninety-one percent but require substantially more data and computational resources. The highest reported accuracies in this range, around ninety-one percent, appear to represent a performance ceiling for current approaches when evaluated on standard datasets.

The proposed hybrid ensemble framework combining Logistic Regression, Random Forest, and XGBoost achieved exceptional performance with ninety-six point nine six percent accuracy, substantially exceeding all previously reported approaches in the comparative analysis. This performance improvement can be attributed to the strategic combination of classifiers with complementary strengths, careful preprocessing and feature engineering, effective handling of class imbalance through SMOTE, rigorous hyperparameter optimization, and nuanced integration of classifier predictions through soft voting. Precision of ninety-six point eight percent, recall of ninety-seven point one percent, and F1-score of ninety-six point nine percent confirm that the model maintains excellent balance between different error types. ROC-AUC of zero point nine nine four and PR-AUC of zero point nine nine two demonstrate excellent discriminative capability. Five-fold cross-validation with standard deviation of zero point zero one two confirms robust performance across different data partitions.

However, rigorous analysis also revealed critical caveats that temper these impressive results and highlight important directions for future work. The exceptional accuracy raises legitimate concerns about potential overfitting, and cross-domain validation experiments confirmed significant performance degradation to seventy-eight point three percent on COVID-19 misinformation datasets. This finding aligns with the work of Hoy and Koulouri and underscores the importance of evaluating models on diverse datasets representing different domains and misinformation types. Computational requirements three point two times higher than individual classifiers may limit deployment in resource-constrained environments. Ensemble complexity reduces interpretability compared to individual models, creating challenges for applications requiring explainability. Adversarial robustness testing revealed vulnerability to simple text manipulations, with accuracy dropping to eighty-two point four percent under paraphrasing attacks.

These findings have significant implications for both researchers and practitioners. For researchers, the study highlights the importance of rigorous validation protocols including cross-domain testing and adversarial evaluation, the value of ensemble methods that leverage complementary classifier strengths, and the critical need for interpretability in high-stakes applications. The substantial performance gap between in-domain and cross-domain evaluation suggests that many published results may overstate real-world applicability, and future work must prioritize generalization assessment. For practitioners, the study provides evidence-based guidance for selecting and deploying fake news detection systems based on specific requirements including accuracy needs, computational constraints, and explainability requirements. The proposed framework offers a practical solution for organizations seeking to deploy fake news detection capabilities while balancing performance with interpretability.

The future of fake news detection lies in the development of systems that are not only accurate but also generalizable, interpretable, robust, and efficient. Cross-domain generalization must be addressed through domain adaptation, transfer learning, and the development of foundation models trained on diverse corpora. Interpretability must be enhanced through improved explanation methods and inherently interpretable architectures. Adversarial robustness must be systematically evaluated and improved through adversarial training and certified defenses. Computational efficiency must be optimized through model compression and hardware acceleration. Multimodal signals must be integrated to provide richer context and enable earlier detection.

The emergence of large language models capable of generating convincing fake content presents both challenges and opportunities for the field. These models can be used to generate increasingly sophisticated misinformation that evades current detection systems, necessitating continuous adaptation and improvement of detection methods. However, the same models may also be leveraged for detection, providing powerful representations of text that capture subtle stylistic and semantic patterns. The development of detection methods that can keep pace with increasingly sophisticated generation techniques represents an ongoing arms race requiring sustained research investment.

Ultimately, automated fake news detection will never achieve perfect accuracy, and the most effective solutions will combine machine capabilities with human judgment. Automated systems can scale to flag potentially problematic content for human review, but human fact-checkers remain essential for making final determinations, particularly in ambiguous or high-stakes cases. The design of effective human-AI collaboration systems, with intuitive interfaces, meaningful explanations, and efficient feedback loops, is essential for real-world impact. Interdisciplinary collaboration between machine learning researchers, human-computer interaction experts, journalists, fact-checkers, and policymakers is necessary to develop systems that are technically sound, practically useful, and socially responsible.

This study contributes to this ongoing effort by providing systematic analysis of current approaches, demonstrating the potential of carefully designed ensemble methods, critically examining fundamental challenges, and outlining comprehensive directions for future research. The proposed framework achieving ninety-six point nine six percent accuracy represents a significant advance, while the accompanying caveats and mitigation strategies provide realistic guidance for deployment. As misinformation continues to evolve in sophistication and impact, the development of robust, interpretable, and deployable detection systems remains one of the most important challenges facing the machine learning community.

REFERENCES

Al Obaidi, S. A., & Çağlıkantar, T. (2024). Automated fake news detection system. *Iraqi Journal for Computer Science and Mathematics*, 5(4). https://doi.org/10.52866/27887421.1200

Al-Tarawneh, M. A. B., Al-Khresheh, A., Al-Irr, O., Kulaglic, A., Danach, K., Kanj, H., et al. (2025). Towards accurate fake news detection: Evaluating ensemble and classical methods. *European Journal of Pure and Applied Mathematics*, 18(2), 6087. https://doi.org/10.29020/nybg.ejpam.v18i2.6087

Dev, D. G., & Bhatnagar, V. (2024). Hybrid RFSVM: Hybridization of SVM and random forest models for detection of fake news. *Algorithms*, 17(10), 459. https://doi.org/10.3390/a17100459

Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, 9(10), e20382. https://doi.org/10.1016/j.heliyon.2023.e20382

Hoy, N., & Koulouri, T. (2025). An exploration of features to improve the generalisability of fake news detection models. *Expert Systems with Applications*, 275, 126949. https://doi.org/10.1016/j.eswa.2025.126949

Ilyas, M. A., Rehman, A., Abbas, A., Kim, D., Naseem, M. T., & Min Allah, N. (2024). Fake news detection on social media using ensemble classifiers. *Computers, Materials and Continua*, 81(3), 4525-4549. https://doi.org/10.32604/cmc.2024.056291

Janssen, J. (2023). Comparative analysis of machine learning algorithms for fake news detection. *Unpublished manuscript*.

Lakshmi, V. D., & Kumari, C. S. (2022). Detection of fake news using machine learning models. *International Journal of Computer Applications*, 183(47), 22-27. https://doi.org/10.5120/ijca2022921874

Mishra, A., Khan, M. H., Khan, W., Khan, M. Z., & Srivastava, N. K. (2022). A comparative study on data mining approach using machine learning techniques: Prediction perspective. In *Pervasive Healthcare* (pp. 153-165). Springer.

Parveen, N., & Khan, M. W. (2024). Proposed algorithm and models for sentiment analysis and opinion mining using web data. *Nanotechnology Perceptions*, 20(6), 3900-3910.

Saini, P., & Khatarkar, V. (2023). A review on fake news detection using machine learning. *Smart Moves Journal IJOScience*. https://doi.org/10.24113/ijoscience.v9i2.511

SK, S., Allada, V. R., Asif, M., Sabeel, U. R., & Shariq, M. B. (2024). Fake news detection using deep learning (LSTM). *International Journal of Creative Research Thoughts*, 12(5), 2320-2882.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559