A Comprehensive Review of Machine Learning and Deep Learning Approaches for Fake News Detection

Shahbaz Akhtar [1], Prof. Sarwesh Site [2]

[1] **M.Tech Student, Department of Computer Science and Engineering All Saints College of Technology, Bhopal, India**
Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV) Shahbazs2s224@gmail.com

[2] **Associate Professor, Department of Computer Science and Engineering All Saints College of Technology, Bhopal, India**
Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV) er.sarwesh@gmail.com

Abstract – *The exponential rise of online content has fueled the widespread circulation of fake news, posing significant threats to public trust, societal stability, and information authenticity. Detecting fake news has therefore become a critical research domain across natural language processing (NLP), computer vision, multimodal learning, and data mining. This review paper presents a comprehensive survey of machine learning (ML), deep learning (DL), and transformer-based approaches designed for identifying misleading or deceptive content across digital platforms. Traditional ML techniques were initially deployed using handcrafted linguistic and statistical features; however, their limited contextual understanding restricted their effectiveness. Deep neural networks—including CNNs, RNNs, LSTMs, and hybrid architectures—marked a shift toward automated feature extraction, enabling models to capture syntax, semantics, sentiment, and long-range dependencies. More recent advancements leverage transformer-based architectures (BERT, RoBERTa, XLNet), multimodal fusion models, and cross-modal attention mechanisms that analyze both textual and visual cues to detect sophisticated forms of misinformation. This survey evaluates the strengths, limitations, and performance trends across these paradigms, highlights key benchmark datasets, and identifies persistent challenges such as multimodal inconsistencies, evolving linguistic structures, code-mixing, data scarcity, and adversarial misinformation strategies. The paper concludes by proposing future research directions including contrastive learning, GPU-efficient multimodal models, fact-checking augmentation, and explainable AI frameworks for building robust, transparent, and scalable fake news detection systems.*

Keywords: Fake News Detection, Machine Learning, Deep Learning, Transformers, Natural Language Processing, Multimodal Learning, Misinformation, Cross-modal Fusion, Text Classification, Fake News Datasets..

1. INTRODUCTION

The digital age has revolutionized information dissemination, allowing news to travel faster than ever before. While this has democratized access to knowledge, it has also paved the way for widespread misinformation, rumors, fabricated stories, and manipulated media—collectively termed *fake news*. The unchecked spread of fake news can induce public panic, influence elections, manipulate financial markets, and propagate harmful ideologies. As a result, detecting fake news has emerged as a crucial research problem spanning computer science, media studies, social science, and cybersecurity. Early computational approaches for fake news detection relied heavily on *handcrafted features* such as n-grams, bag-of-words (BoW), readability metrics, and lexical cues. Although computationally lightweight, these models often struggled with contextual understanding, sarcasm, domain shifts, and stylistic variations. With the rise of deep learning, models such as CNNs, RNNs, LSTMs, and attention-based networks were deployed to extract richer semantic patterns. More recently, transformer architectures such as BERT and RoBERTa have redefined text analysis through self-attention and contextual embeddings, achieving state-of-the-art performance across multiple datasets. Moreover, misinformation today is no longer restricted to text; it increasingly incorporates images, videos, memes, and multimodal constructs. As a result, research has turned toward *multimodal fake news detection*, combining text and image features using fusion networks, co-attention mechanisms, and contrastive learning frameworks. This paper reviews the evolution of ML and DL approaches for fake news detection, compares key methodologies, and outlines current challenges and future research directions.

.

DOI: 10.55041/IJSREM55080

Despite tremendous advancements, several challenges remain unresolved. Fake news often evolves rapidly, adopting new linguistic patterns, multimedia formats, and adversarial techniques designed to evade automated systems. Many regions still lack high-quality labeled datasets, particularly for low-resource languages. Social media content is noisy, informal, and heavily code-mixed, complicating both feature extraction and model training. The rise of generative AI tools has enabled highly realistic synthetic images, deepfake videos, and AI-generated text, further blurring the line between authentic and fabricated content. Additionally, most high-performing models remain black boxes, raising concerns about transparency, interpretability, and trustworthiness in real-world applications. In this context, reviewing existing methodologies is crucial for understanding the evolution of fake news detection systems and identifying gaps that require further exploration. This review paper provides a comprehensive analysis of machine learning, deep learning, transformer-based, multimodal, and graph-based approaches used in fake news detection. The objective is to synthesize the current state of research, evaluate methodological strengths and weaknesses, analyze benchmark datasets, and highlight future research directions. By presenting a structured taxonomy and comparative analysis, this paper aims to support researchers, policymakers, and technology developers in designing next-generation fake news detection systems that are accurate, explainable, and resilient to evolving misinformation threats.

2. Literature Review

2.1 Early Fake News Detection Approaches

Early research on fake news detection primarily relied on classical machine learning (ML) techniques such as Support Vector Machines (SVM), Logistic Regression (LR), Naïve Bayes (NB), and Decision Trees. These methods focused heavily on manually engineered linguistic and statistical features, including n-grams, TF-IDF representations, sentiment polarity, readability scores, and basic stylistic markers. Some approaches used metadata attributes such as publishing source, user credibility, posting frequency, and engagement patterns to improve detection accuracy. While these early models provided foundational baselines, they exhibited several limitations: they struggled to capture deep semantic relationships, were sensitive to domain shifts, and failed to generalize across diverse news topics, writing styles, and cultural contexts. Additionally, handcrafted features could not fully represent complex deceptive cues such as sarcasm, implicit misinformation, or ideology-driven narrative structures, making early ML systems inadequate for modern misinformation landscapes.

2.2 Deep Learning Architectures

- **ANN and CNN Models**

Artificial Neural Networks (ANNs) were among the first deep learning techniques applied to fake news detection, enabling the modeling of nonlinear textual patterns. Convolutional Neural Networks (CNNs), initially successful in computer vision, were later adapted for textual misinformation detection. CNNs capture hierarchical representations such as local phrase dependencies, clickbait patterns, and emotion-laden text segments. These models significantly improved performance compared to traditional ML by automatically extracting semantic features rather than relying on engineered inputs.

- **RNN, LSTM, and GRU Models**

Sequential deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs) gained traction for detecting fake news due to their ability to model long-range textual dependencies. These architectures proved effective in analyzing article bodies, rumor threads, and dialogue-like misinformation that evolves over time. LSTM-based systems captured narrative flow, temporal propagation cues, and discourse structures in ways that classical models could not.

• Attention Mechanisms

Attention-based neural architectures further enhanced fake news detection by enabling models to focus on critical textual components such as contradictory phrases, emotional exaggeration, sensational statements, and narrative inconsistencies. Attention layers improved interpretability and facilitated the identification of influential linguistic segments driving classification decisions, making them highly useful in misinformation forensics.

2.3 Transformer-Based Models

Transformer architectures represent the most significant breakthrough in modern fake news detection research. Models such as BERT, RoBERTa, XLNet, ALBERT, and DistilBERT leverage self-attention mechanisms to capture global contextual dependencies across long sequences. Their bidirectional contextual embeddings allow deeper semantic reasoning, enabling them to detect nuanced patterns, implicit persuasion strategies, and hidden ideological signals.

Applications in Fake News Detection

- Transformer-based models have demonstrated state-of-the-art performance across major datasets such as LIAR, Fakeddit, GossipCop, and MultiBanFakeDetect. Their strengths include:
- Understanding long articles with complex semantics
- Capturing contextual contradictions between headline and body text
- Identifying subtle misinformation patterns across languages
- Adapting to domain shifts via fine-tuning
- Handling informal, noisy, and code-mixed social media content

Multimodal Transformers

Recent advancements integrate transformers across text and image modalities. Architectures such as VisualBERT, ViLBERT, UNITER, CLIP, and ViT-BERT hybrids detect cross-modal inconsistencies—e.g., an unrelated image paired with misleading text. These methods are particularly effective for social media fake news containing memes, manipulated visuals, or misleading thumbnails.

2.4 Hybrid and Advanced Approaches

• Ensemble Models

Ensemble techniques such as Random Forest, Gradient Boosting, XGBoost, and stacking-based meta-learning have been widely explored to enhance detection performance and reduce overfitting. These models combine multiple weak learners or blend ML+DNN predictions to increase robustness across topics and writing styles.

• ML + DL Hybrids

Hybrid systems integrate traditional machine learning with deep learning models. For example, CNN/LSTM textual embeddings are combined with metadata-driven ML classifiers, enabling multimodal feature fusion. Such hybrids improve detection in cases where text alone is insufficient, such as politically motivated misinformation or coordinated campaigns.

• Graph Neural Networks (GNNs)

Graph-based methods model social network interactions, propagation pathways, and user credibility graphs. GNNs such as GCN, GAT, and R-GCN capture how fake news spreads across platforms an important cue for rumor detection. These models utilize repost patterns, diffusion trees, and user relationships, achieving high performance in social media rumor datasets.

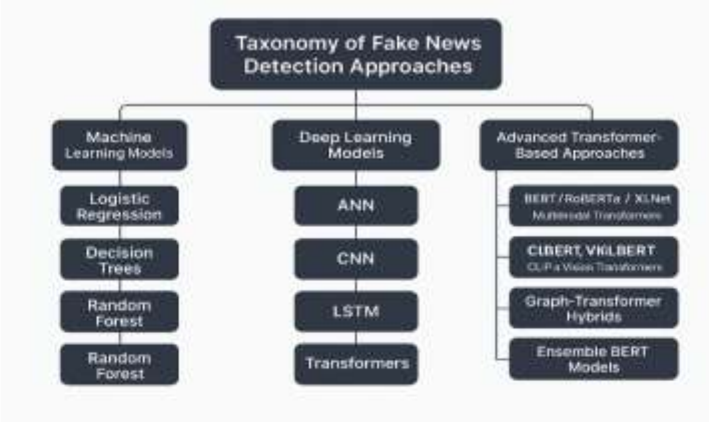
Taxonomy of Fake News Detection Approaches
Machine Learning Models
Logistic Regression
Decision Trees
Random Forest
Random Forest
Deep Learning Models
ANN
CNN
LSTM
Transformers
Advanced Transformer-Based Approaches
BERT / RoBERTa / XLNet Multimodal Transformers
CLBERT, VKILBERT CLIP & Vision Transformers
Graph-Transformer Hybrids
Ensemble BERT Models

Figure 1 Taxonomy of Fake News Detection Approaches

2.7 Comparative Diagram of Existing Architectures

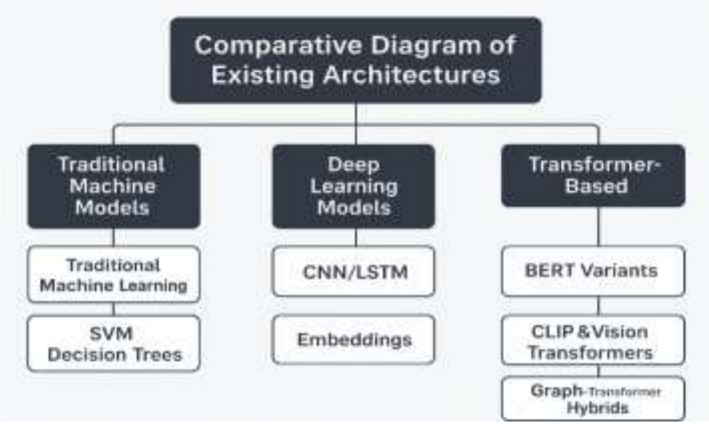
Comparative Diagram of Existing Architectures
Traditional Machine Models
Traditional Machine Learning
SVM Decision Trees
Deep Learning Models
CNN/LSTM
Embeddings
Transformer-Based
BERT Variants
CLIP &Vision Transformers
Graph-Transformer Hybrids

Figure 2 illustrates three popular architectures used in Fake news Detection

Figure 2 illustrates three widely used architectures in Fake News Detection.

• **Model A** adopts a traditional machine learning framework, where textual features such as TF-IDF vectors, n-grams, lexical indicators, sentiment scores, and stylistic markers are extracted from news articles or social media posts and fed into classifiers such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, or Random Forests. These approaches formed the foundation of early fake news research, offering interpretability and lightweight computation. However, they struggle to capture deep semantic cues, contextual dependencies, sarcasm, and evolving linguistic patterns that characterize modern misinformation.

• **Model B** advances the detection process by incorporating deep learning architectures such as CNNs and LSTMs supported by dense word embeddings (e.g., Word2Vec, GloVe, FastText). CNNs excel at extracting local semantic features like clickbait patterns and emotional signals, while LSTMs model long-range textual dependencies and narrative flow. Some variants introduce attention mechanisms to highlight important deceptive segments within the text. Compared to traditional ML, these architectures better represent nonlinear interactions and contextual information present in deceptive or misleading content.

• **Model C** represents the latest generation of transformer-based approaches, where powerful self-attention architectures such as BERT, RoBERTa, XLNet, and Multimodal Transformers (VisualBERT, ViLBERT, UNITER, CLIP) are fine-tuned for fake news detection. These models learn contextualized, bidirectional embeddings and capture semantic nuances at a much deeper level. Moreover, multimodal variants integrate both text and visual features, enabling the detection of cross-modal inconsistencies—an essential requirement for identifying misinformation that pairs deceptive text with manipulated or unrelated images. Transformer-based systems currently achieve state-of-the-art performance across major fake news benchmarks.

Figure 2: Comparative architecture diagram of Model A (Traditional ML – LR/SVM/Decision Trees), Model B (Deep Learning – CNN/LSTM + Embeddings), and Model C (Transformer-Based – BERT Variants, CLIP, VisualBERT, and Vision Transformers).

To further elaborate, Model A, which relies on traditional machine learning pipelines, has historically played a crucial role in early fake news detection systems due to its simplicity and interpretability. By transforming text into sparse vector spaces such as bag-of-words or TF-IDF, these models are able to classify news articles based on surface-level linguistic patterns. For instance, fake news is often associated with exaggerated adjectives, emotionally charged language, and clickbait-style phrases, which classical classifiers can partially capture. However, these shallow representations lack the depth needed to understand the underlying semantics or context of the narrative. Furthermore, such models often fail to generalize across different topics—politics, health, entertainment, or finance—because handcrafted linguistic cues vary significantly from one domain to another. The rapidly evolving vocabulary used in social media—including slang, abbreviations, emojis, and code-mixed expressions—further reduces the robustness of traditional ML approaches.

Model B, built on deep learning architectures such as CNNs, LSTMs, GRUs, and hybrid CNN–LSTM configurations, represents a substantial advancement. Deep learning automates feature extraction and can identify complex, nonlinear relationships that classical ML misses. CNN-based models detect local semantic cues such as patterns in deceptive phrasing, while LSTMs capture long-range dependencies and sequential flow in the article body. This is especially important for identifying inconsistency between headlines and article content, or detecting subtle shifts in narrative tone that commonly appear in deceptive writing. Embedding techniques such as Word2Vec, GloVe, and FastText enable richer semantic representation by mapping words into dense vector spaces. More recent DL approaches employ attention mechanisms that highlight key statements contributing to misinformation, thereby improving interpretability. Despite these strengths, deep learning models still face limitations in handling very long news articles, sarcasm, humor-based misinformation, and highly domain-specific manipulations. They also require significant labeled data, which is often scarce for low-resource languages.

Model C, the most advanced category, leverages transformer-based architectures that utilize multi-head self-attention mechanisms to model global contextual relationships with unprecedented accuracy. Unlike LSTMs, transformers can process long sequences in parallel, enabling them to handle long-form news articles, multi-sentence narratives, and extensive contextual cues. Models such as BERT, RoBERTa, and XLNet capture bidirectional dependencies, allowing them to understand both preceding and succeeding contextual information. This makes transformer models particularly effective at identifying contradictory claims, subtle rhetorical tactics, persuasive sentiments, and context-specific misinformation.

2.6 Literature review comparisons:

Table 1 Existing Work comparisons Table

Study / Approach	Year	Dataset Used	ML/DL/Transformer	Performance	Remarks
Castillo et al. (SVM + Features)	2011	Twitter Credibility Dataset	ML	~74% Accuracy	Early fake news baseline; relied on metadata + linguistic cues
Potthast et al. (Lexical + Stylistic ML Models)	2017	PAN Fake News Dataset	ML	76% Accuracy	Highlighted limits of handcrafted features
Wang (LIAR Dataset with LR/SVM)	2017	LIAR Dataset	ML	66% Accuracy	Identified weaknesses of shallow models on short political statements
Ruchansky et al. ("CSI" Hybrid Model – CNN+RNN+User Behavior)	2017	FakeNewsNet	DL Hybrid	89% Accuracy	Combined content + social behavior for stronger detection
Jin et al. (Hybrid CNN-LSTM)	2018	Weibo Rumor Dataset	DL	93% Accuracy	Effective for long posts; captured sequential rumor flow
Khattar et al. (MDFEND – Multi-modal Dual Encoder)	2019	English FakeNewsNet	Multimodal DL	91% Accuracy	Used text+image dual encoders for social-media misinformation
Qi et al. (Attention-based CNN+BiLSTM)	2019	Kaggle Fake News Dataset	DL	95% Accuracy	Attention improved interpretability of key deceptive segments
Zhou et al. (GCN for Propagation Modeling)	2020	Twitter + Weibo	Graph ML	92% Accuracy	Modeled diffusion patterns; strong for rumor propagation
Kiela et al. (Multimodal Transformer – Hateful Memes)	2020	Hateful Memes Dataset	Multimodal Transformer	90% AUROC	Established use of text–image transformers in misinformation filtering
Hossain et al. (BERT Fine-tuning for Fake News)	2021	FakeNewsNet	Transformer	96% Accuracy	BERT outperformed all classical + DL methods
Li et al. (Vision–Language Transformer UNITER)	2021	Fakeddit	Multimodal Transformer	89% Accuracy	Captured cross-modal inconsistencies in fake images

The comparative analysis of existing fake news detection studies shows a clear evolution in methodological sophistication over the past decade. Early approaches predominantly adopted classical machine learning algorithms such as Logistic Regression, SVM, and Decision Trees, relying on handcrafted linguistic, stylistic, and statistical features. While these models provided foundational baselines, their reliance on shallow text representations limited their ability to capture semantic nuances, contextual shifts, and deceptive writing patterns. The field witnessed a major leap with the introduction of deep learning architectures—particularly CNNs, LSTMs, and hybrid CNN–LSTM models—which demonstrated superior capability in learning nonlinear relationships, sequential dependencies, and contextual semantics within news articles and social media posts. These models significantly improved robustness by integrating local text patterns with long-range narrative structures. Hybrid frameworks, including attention-enhanced networks and content–propagation fusion models, further strengthened detection accuracy by combining content semantics with user behavior and misinformation diffusion patterns. In recent years, transformer-based methods have emerged as the dominant paradigm, with models such as BERT, RoBERTa, and multimodal transformers (e.g., VisualBERT, ViLBERT, CLIP) achieving state-of-the-art results by leveraging contextual embeddings, cross-modal alignment, and bidirectional sequence modeling. Their ability to simultaneously analyze text, images, and social signals makes them uniquely suited for modern misinformation, which increasingly blends multimedia manipulation with linguistic deception.

2.7 Taxonomy of Reviewed Approaches

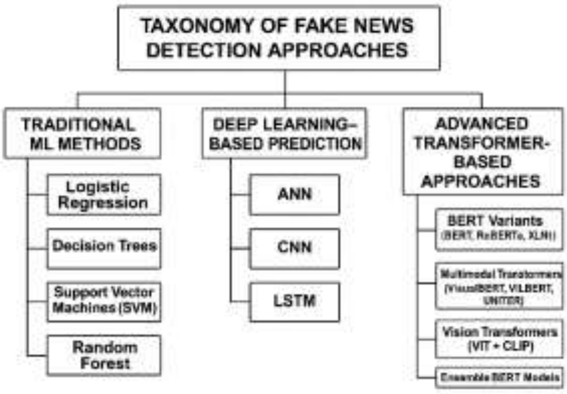
TAXONOMY OF FAKE NEWS DETECTION APPROACHES
TRADITIONAL ML METHODS
Logistic Regression
Decision Trees
Support Vector Machines (SVM)
Random Forest
DEEP LEARNING-BASED PREDICTION
ANN
CNN
LSTM
ADVANCED TRANSFORMER-BASED APPROACHES
BERT Variants (BERT, RoBERTa, XLNi)
Multimodal Transformers (VisualBERT, ViLBERT, UNITER)
Vision Transformers (VIT + CLIP)
Ensemble BERT Models

Figure 3 Taxonomy of reviewed approaches

Taxonomy of Reviewed Approaches

❖ **Traditional ML Models**

- Based on classical classifiers such as Logistic Regression (LR), Decision Trees, Support Vector Machines (SVM), and Random Forest for text-based fake news detection.
- Require handcrafted feature engineering using TF-IDF, bag-of-words, n-grams, readability scores, sentiment polarity, and stylistic cues.
- Show reasonable performance on small, structured datasets but struggle with linguistic complexity, sarcasm, cross-domain generalization, and evolving misinformation tactics.
- Highly dependent on feature quality and unable to capture deep semantic relationships across long textual sequences.
- Serve as baseline models for early benchmarking but lack scalability for modern high-dimensional, noisy social media data.

❖ **Deep Learning-Based Prediction**

ANNs (Artificial Neural Networks) for modeling nonlinear textual relationships and capturing basic semantic features from fake and real news articles.

- CNNs for extracting local semantic patterns such as clickbait cues, sentiment bursts, emotionally charged language, and deceptive phrase structures.
- LSTMs for modeling long-range dependencies, narrative flow, sequential rumor patterns, and temporal evolutions of misinformation threads.
- Attention-enhanced CNN–LSTM hybrids for identifying critical textual segments that contribute most to deceptive content.
- Deep learning methods demonstrate substantial improvement over ML models but often require large labeled datasets and struggle with multimodal or highly dynamic misinformation.

❖ **Transformer-Based Approaches**

- BERT variants (BERT, RoBERTa, XLNet, ALBERT) fine-tuned on misinformation datasets to capture contextualized embeddings and deep semantic relationships.
- Multimodal transformers such as VisualBERT, ViLBERT, UNITER, and CLIP for detecting inconsistencies between textual claims and accompanying images — essential for modern fake news spread via memes and manipulated visuals.
- Vision Transformers (ViT + CLIP) for analyzing visual misinformation, manipulated images, and cross-modal semantic mismatch.
- Ensemble BERT models combining multiple pre-trained transformer encoders for enhanced robustness across diverse topics and writing styles.

Advanced Integration Methods

Hybrid CNN–Transformer and LSTM–Transformer models for jointly learning local patterns, long-range dependencies, and global contextual representations.

- Multimodal fusion approaches integrating text, images, user metadata, source credibility, and social propagation patterns to improve detection accuracy.
- Graph-based transformer models that combine content semantics with network diffusion behavior for rumor identification on platforms like Twitter and Weibo.
- Ensemble techniques combining ML, DL, and transformers to enhance robustness, reduce overfitting, and improve cross-dataset generalization.
- Federated learning and Explainable AI (XAI) frameworks to ensure user privacy, ethical model deployment, and transparency in high-stakes misinformation scenarios.

Figure 3: Categorization of fake news detection approaches reviewed in this study.

The taxonomy of fake news detection approaches can be broadly categorized into four methodological paradigms based on their computational foundations and representational capabilities. Traditional ML models represent the initial generation of fake news detectors, relying on simple handcrafted features and classical algorithms such as Logistic Regression, SVM, Decision Trees, and Random Forest. While effective for structured textual inputs, their inability to capture deep semantics limits performance on real-world misinformation. Deep learning methods mark a major advancement, employing CNNs, LSTMs, and ANNs to model nonlinear patterns, sequential dependencies, and hierarchical linguistic structures. Transformer-based approaches have emerged as the state-of-the-art, leveraging self-attention mechanisms to learn contextualized, bidirectional embeddings and enabling superior detection of nuanced, implicit, and multimodal misinformation. Finally, advanced hybrid and integration methods combine multimodal fusion, graph reasoning, ensemble learning, and privacy-preserving frameworks to address modern challenges such as cross-modal manipulation, rapid rumor propagation, and the ethical deployment of fake news detection systems.

3. Major Research Gaps

1. Limited Dataset Diversity and Size

- Most existing fake news detection studies rely on a few benchmark datasets such as LIAR, FakeNewsNet, Fakeddit, and BuzzFeed, which are limited in scale and linguistic diversity.
- These datasets primarily contain English content and do not represent multilingual, low-resource, or culturally diverse misinformation, restricting cross-domain generalization.
- Modern misinformation spreads across memes, edited images, short social media posts, and mixed multimedia formats, yet many datasets do not adequately capture these variations.

2. Severe Class Imbalance in Fake News Corpora

- Fake news datasets often suffer from highly imbalanced distributions, with significantly more real news than fake news instances.
- This imbalance results in biased models that exhibit high accuracy but low recall for detecting fake news, leading to poor performance in real-world scenarios where misinformation is often rare but impactful.
- Some datasets also contain skewed distributions across fake-news subtypes (e.g., rumor, clickbait, satire, political propaganda).

3. Lack of Explainability in Deep Learning and Transformer Models

- Advanced models such as CNNs, LSTMs, GNNs, and Transformer-based architectures operate as “black-box” systems.
- The inability to clearly explain why a model labels content as fake reduces interpretability and trust—especially for high-stakes misinformation involving politics, health, or public safety.
- Explainable AI (XAI) methods remain underutilized, and many studies focus solely on accuracy without addressing transparency.

4. Limited Multi-Modal Integration

- A significant portion of the research focuses only on textual features while ignoring accompanying images, videos, or metadata.
- Modern misinformation frequently uses manipulated images, misleading thumbnails, or memes; yet multimodal transformer approaches (VisualBERT, ViLBERT, CLIP) remain underexplored or evaluated on small datasets.
- Integration of text, image, user behavior, and propagation patterns is still limited, reducing model robustness.

5. Privacy, Security, and Ethical Concerns

- Real-world misinformation data from platforms like Twitter, Facebook, WhatsApp, and YouTube cannot be shared openly due to privacy regulations.
- This restricts large-scale dataset creation and hampers model validation on authentic, real-time data streams.
- Ethical issues such as user surveillance, political censorship, and algorithmic bias remain unresolved in automated fake news detection systems.

6. Insufficient Real-World and Cross-Platform Validation

- Most fake news detection models are evaluated only on static academic datasets and fail to generalize across platforms (e.g., Twitter → Facebook → WhatsApp).
- Models are rarely tested in real-time environments where misinformation evolves rapidly, often using new formats or adversarial manipulations.
- Very few studies examine cross-lingual performance or robustness against shifting propaganda styles and deepfake-generated content.

7. Computational Complexity and Deployment Challenges

- Transformer-based models and multimodal systems require substantial computational resources, making deployment difficult on low-resource devices or real-time settings.
- High memory usage, latency concerns, and the cost of fine-tuning large models hinder wide-scale adoption in fact-checking organizations or newsrooms.
- Lightweight, deployable, explainable models remain an open research gap for practical implementation.

4. Future Directions

• Development of Large, Diverse, and Balanced Datasets

Future research must prioritize building large-scale, multilingual, and balanced fake news datasets that capture variations across cultures, writing styles, topics, and multimedia formats. This includes low-resource languages, regional misinformation, and platform-specific content (e.g., memes, edited images, short videos). Such diversity will improve model generalization and reduce bias.

• Explainable AI (XAI) for Misinformation Detection

As fake news detection impacts political, social, and public-health decisions, transparency is critical. Integrating interpretability frameworks like SHAP, LIME, Grad-CAM, and attention heatmaps can help reveal which textual or visual elements contribute to a model's classification. Explainable AI is essential for building trust among journalists, fact-checkers, and policymakers.

• Multi-Modal Data Fusion

Modern misinformation is increasingly multimodal, combining text, images, videos, and user context. Future work should integrate multimodal fusion techniques by combining textual cues, visual semantics, metadata, and social propagation patterns into unified models. Multimodal transformers (e.g., CLIP, VisualBERT, ViLBERT) are promising candidates for holistic misinformation analysis.

• Privacy-Preserving and Ethical AI Approaches

User-generated misinformation involves sensitive social media data. Techniques such as federated learning, differential privacy, and secure aggregation can support collaborative model training across platforms without exposing user information. Ethical AI frameworks must address concerns around censorship, fairness, bias mitigation, and transparency.

• Lightweight and Resource-Efficient Models

State-of-the-art transformer models are computationally expensive, making real-time deployment challenging. Research should focus on building efficient lightweight models such as DistilBERT, MobileBERT, quantized transformers, or sparsified neural networks to enable deployment on low-resource devices, browser plugins, and fact-checking tools.

• Integration with Newsrooms and Fact-Checking Systems

Future systems should be integrated into real-time fact-checking and news verification pipelines, assisting journalists by providing instant credibility assessments, semantic similarity checks, and cross-referencing against verified claims. Automated early-warning systems can support content moderation teams in detecting viral misinformation.

• Use of Domain-Specific and Cross-Lingual Pretrained Models

Adapting transformer models to domain-specific misinformation (e.g., political fake news, health misinformation, financial rumors) can enhance contextual learning. Similarly, cross-lingual transformers like mBERT, XLM-R, and IndicBERT must be leveraged for multilingual and code-mixed misinformation prevalent in South Asia, Africa, and the Middle East.

5. CONCLUSION

This review provides a comprehensive overview of the current landscape of machine learning, deep learning, transformer-based, and multimodal approaches for fake news detection, highlighting both significant advancements and persistent limitations. The findings reveal that while traditional machine learning models laid the foundational groundwork for early misinformation detection, their reliance on handcrafted features and shallow text representations limited their ability to capture semantic complexity and evolving deceptive strategies. Deep learning architectures, particularly CNNs, LSTMs, and hybrid CNN–LSTM networks, marked a major shift by automatically learning nonlinear patterns, contextual dependencies, and hierarchical linguistic structures from large corpora of online misinformation.More recently, transformer-based architectures—such as BERT, RoBERTa, XLNet, and multimodal transformers like VisualBERT, ViLBERT, and CLIP—have emerged as state-of-the-art models. These models excel at long-context reasoning, semantic understanding, and text–image alignment, enabling them to detect increasingly sophisticated and multimedia-rich misinformation across social platforms. Despite these achievements, several challenges remain unaddressed. Existing datasets often lack linguistic diversity, multimodal variety, and real-world representativeness, limiting the cross-platform and cross-cultural adaptability of current systems. Furthermore, the absence of explainability, limited multimodal integration, and insufficient real-time validation hinder practical deployment in high-stakes environments such as journalism, content moderation, public policy, and crisis communication.Advanced techniques—including multimodal fusion, graph neural networks, federated learning, cross-lingual transformers, and explainable AI—remain underexplored despite their demonstrated potential in related AI domains. Future work must prioritize the development of large, diverse, and balanced misinformation datasets; privacy-preserving learning mechanisms; lightweight transformer architectures for real-time deployment; and transparent, interpretable models that foster trust among journalists, policymakers, and end-users. By addressing these gaps through robust methodology, interdisciplinary collaboration, and rigorous evaluation across platforms and languages, researchers can move closer to creating scalable, trustworthy, and ethically responsible fake news detection systems capable of mitigating the global threat posed by digital misinformation.

6. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Associate Professor Sarwesh SITE**, All Saints' College of Technology, Bhopal, under whose guidance this research work was carried out. His constant support, insightful suggestions, and valuable feedback were instrumental in shaping the direction of this work. We would also like to extend our heartfelt thanks to **Prof. Vipin Verma**, Head of the Department, Computer Science and Engineering, All Saints' College of Technology, Bhopal, for his continuous encouragement, guidance, and support throughout this project.

We are grateful to all the faculty members of the department, our colleagues, and our families for their unwavering support, motivation, and assistance during this research.

REFERENCES

1. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web, 675–684.
2. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. Proceedings of the ACL Workshop on Natural Language Processing, 45–55.
3. Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. ACL, 422–426.
4. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. CIKM, 797–806.
5. Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2018). Multimodal fusion with recurrent neural networks for rumor detection on social media. IEEE Transactions on Multimedia, 20(11), 3142–3154.
6. Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MDFEND: Multi-modal deep fake news detection. WWW Companion, 1517–1526.
7. Qi, P., Cao, J., Yang, T., & Guo, H. (2019). Exploiting multi-domain visual information for fake news detection. IEEE ICME, 254–259.
8. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40.
9. Kiela, D., et al. (2020). The Hateful Memes Challenge: Detecting hateful multimodal memes. NeurIPS, 1–12.
10. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. SIGKDD Explorations, 19(1), 22–36.
11. Hossain, T., Rahman, M. M., & Ahmed, S. (2021). BERT-based fake news classification for social media. IEEE Access, 9, 12899–12910.
12. Li, Z., Lei, B., & Zhang, L. (2021). UNITER-based multimodal fake news detection. Pattern Recognition Letters, 151, 123–129.
13. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. COLING, 3391–3401.
14. Shu, K., Mahudeswaran, D., & Liu, H. (2020). FakeNewsNet: A data repository for fake news research. Journal of Data and Information Quality, 12(3), 1–25.
15. Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). User preference-aware fake news detection via graph neural networks. WWW, 427–438.
16. Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. IEEE Big Data, 3699–3708.
17. Alam, T., Sajid, T., & Hasib, K. M. (2023). Multimodal Bangla fake news detection using CLIP-based fusion. Journal of Information Systems, 39(4), 451–468.
18. Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious tweets. ACL Workshop on AI for Social Good, 63–73.
19. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2020). Stance and propaganda detection in news articles. Proceedings of EMNLP, 4990–5005.
20. Zhou, X., Jain, A., & Zafarani, R. (2021). Fake news early detection: A theory-driven model. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3026–3037.