Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 14. No. 1, March 2026 Series

Article Citation Format
Ojeniyi, J.A & Kabir, J.M. (2026): Comparative Analysis Of Machine Learning Algorithms For Fake News Detection On Social Media: A Systematic Literature Review. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 14, No. 1. Pp 81-94. www.isteams.net/digitaljournal. dx.doi.org/10.22624/AIMS/DIGITAL/V14N1P6

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 2nd January, 2026
Review Type: Blind Peer
Final Acceptance: 3rd March, 2026

Comparative Analysis of Machine Learning Algorithms for Fake News Detection on Social Media: A Systematic Literature Review

[1]Ojeniyi, J.A & [2]Kabir, J.M.
[1]Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria
[2]Dep of Information Systems and Technology, National Open University of Nigeria, Abuja, Nigeria
E-mails: ojeniyija@futminna.edu.ng, Jabirkabir6@gmail.com
Corresponding E-mail: ojeniyija@futminna.edu.ng

ABSTRACT

The rapid proliferation of fake news across social media platforms constitutes one of the most severe epistemic and societal threats of the digital age. Automated detection using machine learning (ML) and deep learning (DL) techniques has become essential due to the sheer volume, velocity, and viral nature of online misinformation. This systematic literature review (SLR) synthesises 44 peer-reviewed empirical and review studies published between 2020 and 2026, following PRISMA 2020 guidelines, to evaluate the state of ML/DL-based fake news detection on social media. The corpus spans classical ML classifiers (SVM, Random Forest, Naïve Bayes, XGBoost), deep learning architectures (CNN, LSTM, BiLSTM, attention mechanisms), transformer-based models (BERT, CT-BERT, RoBERTa, FakeBERT, ABERT, AraBERT), and advanced ensemble, graph-based, and LLM-augmented approaches that integrate social context, user profiling, knowledge graphs, and contrastive learning. Findings strongly support both tested hypotheses: H_2 – deep learning and transformer models consistently outperform classical ML, with transformer ensembles achieving F1-scores of 93–99.1% across benchmark datasets including LIAR, FakeNewsNet, ISOT, Weibo, and COVID-19 corpora; and H_3 – ensemble and hybrid methods incorporating social and contextual features deliver superior robustness, with the best hybrid systems (TAM-ATOA, MG-CL, KeepUp) outperforming single-model baselines by 3–12 percentage points. Key challenges identified include dataset bias, class imbalance, multilingual limitations, lack of explainability, real-time deployment constraints, and adversarial robustness. The review provides replicable evidence base and identifies actionable directions for the next generation of reliable, ethical, and deployable fake news detection systems.

Keywords: Fake News Detection, Machine Learning, Deep Learning, Transformer Models, NLP, Misinformation, Social Media, BERT, Ensemble Learning, Systematic Literature Review

1. INTRODUCTION

The proliferation of fake news on social media has emerged as one of the most pressing challenges of the digital age. Platforms such as Twitter (now X), Facebook, TikTok, and Instagram lack the editorial oversight of traditional media, enabling misinformation to spread at

unprecedented speed and scale. Empirical evidence confirms that false news reaches 6–20 times more people than truthful content (Vosoughi et al., 2018), with consequences spanning eroded public trust, election interference, COVID-19 vaccine hesitancy, and heightened social polarisation (Lazer et al., 2018; Allcott & Gentzkow, 2017). The 2016 U.S. presidential election demonstrated how coordinated disinformation campaigns exploit algorithmic amplification, while the COVID-19 infodemic which the World Health Organization (WHO) identified as a parallel public health crisis illustrated how false health information can translate directly into preventable deaths. According to Alghamdi et al. (2023), more than 3,500 false claims about COVID-19 were identified within the first two months of the pandemic alone, and over 800 coronavirus-related deaths worldwide in early 2020 were directly attributable to COVID-19 misinformation.

Manual factchecking cannot scale to the billions of daily posts generated on social platforms. Consequently, automated fake news detection (FND) using ML and DL has become a critical research domain (Zhou & Zafarani, 2020). Early approaches relied on classical ML classifiers with hand-crafted features such as TF-IDF, n-grams, and sentiment scores. The deep learning era introduced CNNs, LSTMs, and attention mechanisms capable of capturing richer semantic representations. The transformer revolution culminated in BERT and its variants, which deliver contextualised embeddings from pre-trained models fine-tuned on domain-specific data, achieving F1-scores in the high 90s on benchmark datasets. The most recent generation of approaches incorporates social network graph structures, knowledge bases, user behavior signals, multimodal inputs, and large language model (LLM) augmentation.

Despite substantial progress, persistent tensions exist between benchmark performance and real-world generalisability, between accuracy and explainability and between text only approaches and context aware models. This SLR addresses these tensions by synthesising 44 studies published between 2020 and 2026 and directly testing two core hypotheses derived from the author's master's thesis research:

H_2: Deep learning and transformer-based models consistently outperform classical ML approaches in fake news detection performance.

H_3: Ensemble and hybrid models incorporating social and contextual features achieve superior robustness and accuracy compared to single-model approaches.

By rigorously evaluating these hypotheses across multiple model families, datasets, and languages, this review establishes a replicable evidence base, maps the current landscape of FND research, identifies critical gaps, and proposes actionable directions for future empirical work.

2. BACKGROUND AND THEORETICAL FRAMEWORK

2.1 The Fake News Problem on Social Media

Fake news on social media exhibits characteristics that fundamentally distinguish it from traditional misinformation. Content is often short form (particularly on Twitter/X), laden with emotional language, amplified through algorithmic recommendation systems, and spread by both organic users and automated bots.

Detection must therefore extend beyond textual semantics to incorporate propagation patterns, user credibility, stance consistency, source analysis, and temporal dynamics. The bibliometric analysis by Zeeshan et al. (2025), drawing on 649 publications from the Web of Science database spanning 1991–2023, documents a 14.81% annual growth rate in fake news detection research. Publications peaked sharply in 2021, driven by the COVID-19 infodemic, which catalysed an unprecedented volume of health misinformation across Twitter, Facebook, and WhatsApp. IEEE Access and Expert Systems with Applications dominate as publication venues, while China, the United States, India, Pakistan, and Saudi Arabia contribute the largest national research outputs. This bibliometric landscape confirms that fake news detection has rapidly transitioned from a niche topic to a mainstream research priority.

The taxonomy of false content encompasses deliberate disinformation (intentionally fabricated to deceive), misinformation (false but not intentionally misleading), satire and parody (mimicking journalism for entertainment), and emerging forms such as AI-generated synthetic news. Each category presents distinct detection challenges: political disinformation is often sophisticated and deliberately mimics legitimate reporting; health misinformation may cite real but misinterpreted studies; AI-generated fake news may be linguistically indistinguishable from human-authored real news. Tanaja et al. (2025) note that this categorical diversity requires models capable of detecting multiple distinct patterns of inauthenticity simultaneously.

2.2 Evolution of Detection Approaches

The evolution of fake news detection mirrors the broader progression of NLP and AI. Classical approaches used hand-crafted features combined with traditional classifiers. Multiple studies in this corpus (Wiese & Wiese, 2022; Kamaruddin et al., 2025; Paulin & Balaba, 2025; Suresh et al., 2024; Hosea et al., 2023; Twum, 2025) evaluate this paradigm, consistently demonstrating 85–95% accuracy. These classifiers are interpretable, computationally efficient, and trainable on modest hardware, making them suitable for resource-constrained deployments, but they are fundamentally limited by their inability to capture deep contextual meaning.

The deep learning era introduced neural architectures capable of learning feature representations automatically from data. CNNs excel at extracting local n-gram-like features; LSTMs and their bidirectional variants capture sequential dependencies across document length; attention mechanisms allow selective focus on the most discriminative segments. Hybrid architectures combining CNN with LSTM, or adding attention layers atop recurrent networks, consistently outperform single-architecture models. Das et al. (2025) and Embark (2025) demonstrate this with CNN-RNN hybrids achieving 94–95% accuracy on benchmark datasets; Jin and Wang's (2025) TAM-ATOA framework achieves 88–93% accuracy across four benchmarks using triple-path attention.

The transformer revolution, initiated by the 'Attention Is All You Need' architecture (Vaswani et al., 2017) and operationalised for NLP through BERT (Devlin et al., 2019), transformed the performance ceiling for fake news detection. Domain-specific pre-training (CT-BERT on 160M COVID-19 tweets; AraBERT on Arabic corpora; multilingual BERT for cross-lingual transfer) further elevated performance. The most recent frontier incorporates LLMs for data augmentation (Arık et al., 2026), graph neural networks for propagation modelling (Huang & Liu, 2026), and unified frameworks combining knowledge extraction with social platform analysis (Wasim et al., 2026).

3. METHODOLOGY

3.1 Study Design and Protocol

This systematic literature review follows PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines as adapted for computer science research. The review protocol was developed as part of the author's master's thesis at NOUN and pre-registered in the thesis proposal. The central research question is: How do machine learning and deep learning techniques address fake news detection on social media, which model families achieve superior performance, and what are the key methodological tensions and research gaps?

3.2 Search Strategy

Systematic searches were conducted across five databases: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar. Boolean operator queries combined three conceptual domains: (1) content terms ('fake news detection' OR 'misinformation detection' OR 'disinformation' OR 'infodemic' OR 'rumour detection'); (2) technique terms ('machine learning' OR 'deep learning' OR 'BERT' OR 'transformer' OR 'ensemble' OR 'CNN' OR 'LSTM' OR 'GNN' OR 'LLM'); and (3) context terms ('social media' OR 'Twitter' OR 'Facebook' OR 'online platform' OR 'microblog'). Searches were restricted to January 2020–March 2026 to ensure contemporaneous relevance.

Table 1: Search Terms Applied Across Databases

Category 1: Content	Category 2: Technique	Category 3: Context
"Fake news detection" OR "Misinformation detection" OR "Disinformation" OR "Infodemic" OR "Rumour detection"	"Machine learning" OR "Deep learning" OR "BERT" OR "Transformer" OR "Ensemble" OR "CNN" OR "LSTM" OR "GNN" OR "LLM"	"Social media" OR "Twitter" OR "Facebook" OR "Online platform" OR "Microblog" OR "Weibo"

3.3 Inclusion and Exclusion Criteria

Studies were included if they: (1) focused on fake news, misinformation, or rumour detection as a primary task; (2) explicitly employed ML or DL techniques; (3) targeted social media, online news, or related digital platforms; (4) reported quantitative performance metrics (accuracy, F1, precision, recall, or AUC); and (5) were peer-reviewed journal articles, conference proceedings, or systematic reviews published in English between 2020 and 2026. Studies were excluded if they: focused exclusively on nondigital domains, reported only qualitative findings, were duplicates or earlier versions of updated studies, or were not peer-reviewed. Both empirical studies and systematic reviews were eligible for inclusion.

3.4 Study Selection and Quality Assessment

Initial database searches returned 312 records after deduplication. Title and abstract screening eliminated 184 records clearly not meeting inclusion criteria, yielding 128 full-text articles for detailed review. Applying all inclusion and exclusion criteria resulted in a final corpus of 44 studies. The PRISMA flow records: Identification (n=312) → Screening (n=128) → Eligibility (n=44). Quality assessment evaluated each study on five criteria: clarity of methodology, appropriateness of dataset, reproducibility of results, significance of findings, and identification of limitations. All 44 studies were rated as meeting minimum quality thresholds for inclusion.

3.5 Data Extraction and Analysis

A standardised extraction template captured: author(s) and year; publication venue; study type (empirical/review/comparative); model architecture and family; datasets and languages; feature engineering approach; best reported performance metrics; key findings; identified limitations; and relevance to H_2 and H_3. Extracted data were thematically coded by model family, feature type, and performance tier. Cross-study comparisons employed percentage performance differences where metrics were reported on comparable datasets or tasks.

4. FINDINGS

4.1 Overview of Included Studies

The 44 included studies span 2020–2026, with a pronounced concentration in 2024–2026 (n=31, 70.5%), reflecting the accelerating pace of transformer and LLM research. Publication venues include Procedia Computer Science (n=11), Machine Learning with Applications (n=6), Egyptian Informatics Journal (n=4), Expert Systems with Applications (n=3), Knowledge-Based Systems (n=2), Information Processing & Management (n=2), Journal of King Saud University (n=2), and various specialised journals and conference proceedings covering computer science, information science, and applied NLP. Geographically, studies originate from institutions across more than 20 countries, with significant contributions from China, Australia, Saudi Arabia, India, Turkey, the Philippines, Ethiopia, Indonesia, Nigeria, and the United Kingdom.

Studies were categorised into four methodological families: classical ML (n=9, 20.5%); deep learning and hybrid architectures (n=13, 29.5%); transformer-centred models (n=12, 27.3%); and ensemble, graph-based, and LLM-augmented approaches (n=10, 22.7%). Seven studies are primarily review or survey articles that synthesise existing empirical work. The predominant benchmark datasets are LIAR (Wang, 2017), FakeNewsNet (Shu et al., 2020), ISOT (Ahmed et al., 2018), Weibo21/Weibo20 (Chinese microblogs), PolitiFact, GossipCop, and various COVID-19-specific corpora. A growing minority of studies addresses non-English languages, including Turkish (Arık et al., 2026), Arabic (Yousif, 2025), Amharic (Gemeda Yigezu et al., 2024), Chinese (Huang & Liu, 2026), and Indonesian (Satria et al., 2025a).

Table 2: Master Extraction Table – Selected Included Studies (Representative Sample)

Study (Year)	Model / Approach	Dataset(s)	Best Result	Key Finding	Limitations / H_2 & H_3
Zeeshan et al. (2025)	Bibliometric analysis (R Bibliometrix, VOSviewer)	Web of Science (649 papers, 1991–2023)	N/A (review)	14.81% annual growth; peak 2021 (COVID); DL/transformer dominance in recent literature	English-language bias; no empirical comparison. H_2/H_3: contextual support
Huang & Liu (2026)	MG-CL: 4-graph + GMLP fusion + RoBERTa + contrastive learning	Weibo21, Weibo20, CHECKED (Chinese)	Acc 93.45%, F1 86.85%	Multi-feature graphs > single view; contrastive learning enables 1%	Chinese only; computational cost. H_2: Strong. H_3: Strong

Study (Year)	Model / Approach	Dataset(s)	Best Result	Key Finding	Limitations / H_2 & H_3
Vasist & Sebastian (2022)	Max Voting Ensemble (SVM+DT+LR), AdaBoost, XGBoost, LSTM	ISOT + COVID-19 datasets	Acc 98.4% (Max Voting)	Ensemble outperforms LSTM (92.9%); thematic heterogeneity degrades accuracy	Text-only; English only. H_2: Moderate. H_3: Strong
Jin & Wang (2025)	TAM-ATOA: Triple-Attention + Advanced Tailor Optimisation Algorithm	FakeNewsNet, LIAR, PolitiFact, GossipCop	Acc 93%, F1 93%; AUC-ROC 0.96	3-channel attention outperforms BERT/BiLSTM by 3–5%; ATOA converges 30–50% faster	English only; high computational cost. H_2: Strong. H_3: Moderate
Alghamdi et al. (2023)	CT-BERT+BiGRU; BERT+CNN, BiLSTM, mCNN; RoBERTa	COVID-19 Constraint@AAAI2021 (6,420 tweets)	F1 98.54% (CT-BERT+BiGRU)	Domain-specific pre-training + BiGRU > all baselines; fine-tuning >> feature-extraction	Small dataset; English only; hyperparameter sensitivity. H_2: Very Strong. H_3: Strong
Hosea et al. (2023)	SVM + Word2Vec + Doc2Vec; Lagrangian Duality	Kaggle 26,000-instance dataset	Acc 95.74% (SVM)	SVM+Lagrangian duality outperforms KNN (79%), NB (75%), CSI (89.2%)	No DL comparison; Nigerian context gap. H_2: Partial. H_3: N/A
Arık et al. (2026)	Turkish LLaMA-3 8B augmentation + XGBoost ensemble	Turkish Political Fake News Dataset (TPFND, 9,230 items)	Fake news rate 97.62%	LLM augmentation increases minority class detection from 91.12% to 97.62%	Turkish only; slight precision trade-off. H_2: Indirect. H_3: Strong
Wasim et al. (2026)	KeepUp: Knowledge extraction + social platform engagement + user	Twitter + FakeNewsNet	F1 96.9%	Knowledge + social context > text-only; user profiling adds 4% F1; robust to adversarial inputs	Requires platform API; privacy concerns. H_2: Strong. H_3: Very Strong

Study (Year)	Model / Approach	Dataset(s)	Best Result	Key Finding	Limitations / H_2 & H_3
	profiling				
Alghamdi et al. (2025)	ABERT: Adapted BERT for human/AI-generated fake news detection	Custom human+AI-generated corpus	F1 96.4%	ABERT detects AI-generated fake news; adapted pre-training crucial; general BERT degrades	Emerging problem; AI content evolving rapidly. H_2: Very Strong. H_3: Moderate
Tanaja et al. (2025)	FakeBERT: BERT + style analysis + credibility verification	LIAR, FakeNewsNet	F1 98.2%	Multi-signal integration critical; FakeBERT with style analysis achieves 98.2% F1	Requires credibility DB; English; high inference cost. H_2: Very Strong. H_3: Strong
Reddy et al. (2024)	Ensemble: CNN + LSTM + BERT voting	LIAR, FakeNewsNet, ISOT	F1 96.8%	DL ensemble outperforms individual models by 2–4%; cross-architecture diversity beneficial	English; fixed dataset; no social context. H_2: Strong. H_3: Very Strong
Gemeda Yigezu et al. (2024)	Ethio-Fake: Multilingual BERT + Amharic NLP features + XAI	Ethio-Fake (Amharic corpus)	F1 88.3%	First comprehensive Amharic FND study; XAI highlights cultural nuances	Low-resource; limited dataset. H_2: Strong. H_3: Moderate
Nair et al. (2024)	Knowledge-based DL: knowledge graph embeddings + BERT	Twitter FND dataset	Acc 96.8%	Knowledge graphs ground BERT predictions; KG+BERT > BERT alone by 2.3%	KG completeness limits; English; static knowledge. H_2: Very Strong. H_3: Strong

4.2 Classical Machine Learning Approaches

Nine studies in this corpus employ classical ML as the primary methodology. These studies collectively demonstrate that classical ML remains viable, particularly in resource-constrained or interpretability-demanding contexts, while consistently underperforming relative to deep learning counterparts. SVM emerges as the strongest individual classical classifier across studies, with Hosea et al. (2023) achieving 95.74% accuracy using SVM with Lagrangian Duality on a Kaggle dataset, and Kamaruddin et al. (2025) reporting 92.1% using SVM with TF-IDF features. Random Forest consistently places as the strongest ensemble approach within the classical paradigm, achieving 93.1% (Wiese & Wiese, 2022), 94.5% (Jouhar et al., 2024), 94.1% (Suresh et al., 2024), and 93.8% (Thamiliny et al., 2025) across diverse datasets.

The most instructive study in this category is Vasist and Sebastian (2022), whose systematic comparison of MNB, Passive Aggressive Classifier, Max Voting Ensemble (SVM+DT+LR), Random Forest, AdaBoost, XGBoost, and LSTM across two datasets varying in thematic diversity demonstrates that the Max Voting Ensemble achieves 98.4% accuracy on the mixed dataset – higher than LSTM (92.9%) and all individual classical classifiers. This finding is methodologically significant: it challenges the simplistic narrative that DL always outperforms classical ML, demonstrating that well-designed classical ensembles can match neural networks on specific benchmarks, particularly with abundant and well-structured training data.

Across classical ML studies, the performance ceiling for single classifiers is approximately 92–94% accuracy, rising to 95–98% for optimal ensembles. Persistent limitations include: inability to capture long-range semantic dependencies without explicit feature engineering; sensitivity to class imbalance; poor cross-domain generalisation; and difficulty processing user behavioural or network features without significant preprocessing overhead. Arık et al. (2026) address one major limitation which is class imbalance in the Turkish fake news domain through LLM-based synthetic data augmentation, demonstrating that LLaMA-3 8B-generated synthetic minority-class samples raise fake news detection recall from 91.12% to 97.62% when training XGBoost classifiers.

4.3 Deep Learning and Hybrid Architectures

Thirteen studies evaluate deep learning architectures, with consistent findings supporting H_2. The core finding across this category is that hybrid architectures combining CNN and LSTM/GRU consistently outperform either architecture alone, and that adding attention mechanisms provides a further improvement of 1–3 percentage points. Das et al. (2025) demonstrate this systematically: CNN-only achieves 91.3% accuracy, RNN-only 88.9%, and CNN-RNN hybrid 94.7% – a meaningful gap that illustrates the complementarity of local feature extraction (CNN) and sequential modelling (RNN). Alghamdi et al. (2023) provide the most rigorous architectural comparison in this category, testing 22 model variants combining BERT/CT-BERT as encoders with CNN, LSTM, GRU, BiLSTM, BiGRU, mCNN, and CNN-BiGRU downstream architectures.

Their finding that CT-BERT+BiGRU (98.54% F1) outperforms CT-BERT alone (97.70% F1) by 0.84 points and all individual DL architectures using GloVe by 5–9 points establishes a clear hierarchy: domain-specific PLM > general PLM > hybrid DL > single DL architecture. Jin and Wang (2025) propose the most architecturally novel deep learning system in the corpus: TAM-ATOA, which simultaneously processes three information channels (textual content, contextual metadata, and user behavioural signals) through parallel attention modules fused via a dense residual network with capsule layers.

The Advanced Tailor Optimisation Algorithm (ATOA) achieves convergence 30–50% faster than AdamW with superior final performance across four benchmark datasets (FakeNewsNet: Acc 93%, F1 93%; LIAR: Acc 88%, F1 88%; PolitiFact: Acc 91%; GossipCop: Acc 90%). Bidirectionality is consistently beneficial: BiLSTM outperforms LSTM by 2–4 points, and BiGRU outperforms GRU by similar margins. This is attributable to backward context being informative for fake news detection – deceptive framing often manifests in conclusion sentences that are best understood with prior context already encoded. GRU variants are computationally more efficient than LSTM with comparable performance at smaller dataset sizes, making them preferable for resource-constrained deployments.

4.4 Transformer-Based Models

Twelve studies focus primarily on transformer models, with this category providing the strongest empirical evidence for H_2. BERT and its domain-specific variants consistently achieve F1-scores of 93–99.1% across benchmark datasets, representing the current performance frontier for single-model approaches. The most impactful transformer study in the corpus is Alghamdi et al. (2023), which demonstrates that domain-specific pre-training is the single most powerful lever for transformer performance. CT-BERT – pre-trained on 160 million COVID-19 tweets – consistently outperforms generalist BERT and RoBERTa on COVID-19 fake news detection, while CT-BERT+BiGRU achieves state-of-the-art 98.54% F1. This study definitively establishes three hierarchical principles for transformer deployment: (1) domain-specific pre-training > general pre-training; (2) fine-tuning > feature extraction (frozen weights reduce F1 by up to 29.64 points for CT-BERT); and (3) complex downstream architectures (BiGRU, BiLSTM) > simple classification head.

Alghamdi et al. (2025) address an emerging threat – AI-generated fake news – with ABERT, an adapted BERT model achieving 96.4% F1 in distinguishing human-written from LLM-generated fake news. Tanaja et al. (2025) achieve 98.2% F1 by integrating FakeBERT with style analysis and credibility verification – a three-pronged approach that combines transformer embeddings with stylometric signals and external credibility scores. Yousif (2025) demonstrates multilingual transformer applicability with AraBERT achieving 94.7% F1 on Arabic social media, outperforming Arabic classical ML by 9.4 points. The Khan et al. (2021) benchmark study provides crucial cross-model evidence: BERT achieves 94.1% F1 on LIAR compared to XGBoost (87.2%), SVM (85.7%), and LR (81.3%), demonstrating a consistent 7–13 point advantage across all classical baselines. These benchmark comparisons across multiple studies with multiple datasets constitute the strongest quantitative evidence for H_2, with transformer superiority over classical ML consistent across five or more independent studies.

4.5 Ensemble, Graph-Based, and Advanced Hybrid Approaches

Ten studies represent the most sophisticated methodological tier, providing the strongest evidence for H_3. These approaches combine multiple model families, information modalities, external knowledge sources, or novel learning paradigms to achieve robustness that individual models cannot. Huang and Liu (2026) propose MG-CL, the most architecturally innovative approach in the corpus. Four complementary linguistic graphs – raw text co-occurrence (CO), part-of-speech (POS), named entity (NE), and semantic dependency (SD) – are independently encoded and fused through a Gated MLP (GMLP) mechanism combined with RoBERTa embeddings and a cluster-guided contrastive learning objective. MG-CL achieves 72.61–93.45% accuracy across three Chinese Weibo datasets using only 1–5% of training data.

The ablation study definitively confirms H_3: removing any single graph degrades performance, removing contrastive learning reduces F1 by 2.3 points, and the full ensemble consistently outperforms all ablated variants. Wasim et al. (2026) present KeepUp, which integrates three complementary information sources: knowledge extraction from external knowledge bases, social platform engagement metrics (likes, shares, comment sentiment), and user profiling (account age, verification status, posting history, network centrality). Achieving 96.9% F1 on Twitter and FakeNewsNet, KeepUp exemplifies H_3: no individual component achieves the combined system's performance. Reddy et al. (2024) demonstrate the ensemble principle within the deep learning paradigm by combining CNN, LSTM, and BERT predictions through weighted voting, achieving 96.8% F1 on LIAR, FakeNewsNet, and ISOT – consistently outperforming the best individual model by 2–4 percentage points.

4.6 Multilingual, Low-Resource, and Cross-Domain Findings

A critical gap identified across the corpus is the dominance of English-language research. Of 44 studies, approximately 35 (79.5%) evaluate exclusively on English datasets. Non-English contributions include: Turkish (Arık et al., 2026); Arabic (Yousif, 2025); Amharic/Ethiopic (Gemeda Yigezu et al., 2024); Chinese (Huang & Liu, 2026); and Indonesian (Satria et al., 2025a). Gemeda Yigezu et al. (2024) pioneer fake news detection for Amharic, one of the most under-resourced major African languages, constructing the Ethio-Fake dataset and applying multilingual BERT with Amharic-specific NLP features and Explainable AI (XAI) components. Achieving 88.3% F1, the study demonstrates that multilingual BERT provides a strong foundation even for low-resource languages, but domain-specific fine-tuning data and culturally-sensitive feature engineering are essential.

5. DISCUSSION

5.1 Testing H_2: Deep Learning and Transformer Superiority

The 44 study corpus provides robust, consistent, and cross-dataset support for H_2. Across all studies reporting direct comparisons between transformer models and classical ML, transformer models outperform classical approaches by 7–13 percentage points in F1-score (Khan et al., 2021; Satria et al., 2025b; Alghamdi et al., 2023; Yousif, 2025). The performance advantage is consistent across English, Arabic, Turkish, Chinese, and Indonesian language contexts. Within deep learning, bidirectional architectures consistently outperform unidirectional variants by 2–4 points, and attention mechanisms provide a further 1–3 point improvement over recurrent models without attention. However, the evidence for H_2 is nuanced in three important ways. First, the Vasist and Sebastian (2022) study demonstrates that a well-designed Max Voting Ensemble of classical classifiers (98.4% accuracy) can exceed individual DL models (LSTM: 92.9%) on specific benchmarks, suggesting the H_2 advantage is not universal and depends on dataset characteristics. Second, the performance advantage of transformers is largest on complex semantic tasks and smallest on straightforward binary classification with abundant, balanced training data.

Third, computational efficiency must factor into H_2 evaluation: classical ML inference is orders of magnitude faster, making it preferable for real-time monitoring at social media scale despite lower accuracy. Domain-specific pre-training emerges as the single most powerful transformer optimisation lever. CT-BERT's 98.54% F1 versus generalist RoBERTa's 95.85% on COVID-19 data (Alghamdi et al., 2023), AraBERT's 94.7% versus Arabic classical ML's 85.3% (Yousif, 2025), and ABERT's 96.4% versus standard BERT's lower performance on AI-generated content (Alghamdi et al., 2025) collectively establish that models should be pre-trained on corpus data as close as possible to the target domain, language, and content type.

5.2 Testing H_3: Ensemble and Hybrid Superiority

Evidence for H_3 is equally compelling across all methodological families. Within classical ML, Max Voting Ensembles (SVM+DT+LR) and Random Forest consistently outperform individual classifiers by 3–8 percentage points. Within deep learning, CNN-LSTM hybrids outperform CNN-only by 3–4 points and RNN-only by 5–6 points. Transformer ensembles outperform the best individual transformer by 2–4 points. Cross-paradigm hybrids combining transformers with graph networks (Alghamdi et al., 2024), knowledge bases (Nair et al., 2024; Wasim et al., 2026), social context (Rao et al., 2024; Embark, 2025), or contrastive learning (Huang & Liu, 2026) consistently achieve the highest performance in their respective categories.

The pattern is theoretically coherent: no single model or feature type captures the full complexity of fake news detection. Text conveys semantic content but misses propagation dynamics. Graph networks capture social structure but ignore linguistic nuance. Knowledge graphs provide factual grounding but are incomplete. User behaviour signals capture social dynamics but are manipulable by sophisticated actors. Ensemble and hybrid approaches reconcile these complementary limitations, achieving robustness that individual models cannot. KeepUp (Wasim et al., 2026) is the clearest demonstration: fusing knowledge extraction, social platform engagement, and user profiling achieves 96.9% F1 while simultaneously being more robust to adversarial inputs than text-only counterparts.

5.3 Identified Tensions and Research Gaps

Several persistent tensions emerge from synthesising the 44 studies. The benchmark versus real world tension is pervasive: models achieving 97–99% F1 on curated benchmark datasets may perform significantly worse on raw social media data due to distribution shift, adversarial manipulation, and temporal drift. The accuracy-versus-explainability tension is fundamental to deployment. High-performing transformer and graph models are essentially black boxes, providing no human-interpretable rationale for classification decisions. Only Gemeda Yigezu et al. (2024) incorporate Explainable AI (XAI) techniques as a core component.

Critical research gaps identified across the corpus include: (1) systematic multilingual and cross-lingual evaluation – only 5 of 44 studies address non-English languages; (2) adversarial robustness testing – almost no study evaluates models against adversarially crafted fake news; (3) temporal robustness – performance on time-shifted data after model training is rarely reported; (4) real-time deployment evaluation – most research is confined to offline batch evaluation; (5) AI-generated misinformation detection at scale; (6) ethical and bias analysis; and (7) dataset recency bias – LIAR (2017) and ISOT (2018) remain dominant benchmarks despite being created before the current LLM era.

6. CONCLUSION

This systematic literature review synthesised 44 peer-reviewed studies published between 2020 and 2026, evaluating machine learning and deep learning approaches to fake news detection on social media. The review provides robust, cross-dataset, and multilingual evidence for both tested hypotheses. H_2 is strongly supported: transformer-based models achieve F1-scores of 93–99.1% across benchmark datasets, consistently outperforming classical ML (85–95%) and standard deep learning approaches (91–96%) by margins of 7–13 and 2–3 percentage points respectively.

Domain-specific pre-training, fine-tuning, and complex downstream architectures (BiGRU, BiLSTM) emerge as the most impactful optimisation levers for transformers. H_3 is equally supported: ensemble and hybrid methods consistently outperform their single-model counterparts across all methodological families, with the largest gains achieved by cross-paradigm hybrids combining transformers with graph networks, knowledge bases, social context features, or contrastive learning objectives. The field has progressed from simple TF-IDF classifiers to sophisticated multi-signal architectures that jointly model text, propagation structure, social network dynamics, and factual knowledge. However, the dominant research paradigm – training and evaluating on pre-curated, English-language datasets under laboratory conditions – creates a systematic gap between reported performance and real-world deployability. The most pressing challenges for the next generation of FND research are not primarily technical but operational: how to build models that are interpretable enough for accountable deployment, robust enough to withstand adversarial manipulation, generalised enough to perform across languages and fake news domains, fast enough for real-time social media monitoring, and ethical enough to avoid reproducing or amplifying existing social biases.

Actionable research priorities emerging from this review are: (1) development of multilingual and cross-lingual benchmark datasets and evaluation protocols; (2) systematic adversarial robustness evaluation and adversarially-robust training methods; (3) integration of Explainable AI techniques as standard evaluation criteria alongside accuracy metrics; (4) real-time deployment architecture research with explicit latency and throughput constraints; (5) longitudinal evaluation of model performance drift over time; (6) specialised detection approaches for AI-generated misinformation; (7) ethical frameworks that audit FND models for demographic and political bias; and (8) low-resource and cross-lingual transfer learning methods that extend FND capabilities beyond the dominant English-language paradigm.

The proliferation of fake news constitutes a genuine threat to democratic discourse, public health, and social cohesion. The research synthesised in this review demonstrates that technical solutions with strong benchmark performance exist. Translating these solutions into reliable, accountable, equitable, and deployable real-world detection systems remains the defining challenge for the interdisciplinary field of computational misinformation research.

DECLARATIONS

- **Funding:** Self-funded as part of master's thesis research, National Open University of Nigeria.
- **Competing Interests:** The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.
- **Ethical Approval:** Not applicable (secondary data synthesis only).
- **CRediT Statement:** Kabir Jabir Muhammad: Conceptualisation, Methodology, Formal Analysis, Investigation, Data Curation, Writing (Original Draft), Writing (Review & Editing), Project Administration.

REFERENCES

Abdulsalami, A. O., et al. (2026a). A web-based fake news detection system using hybrid machine learning. Journal of Systematic and Modern Science Research, 11(9), 57–68.

Abdulsalami, A. O., et al. (2026b). A hybrid deep learning and differential evolution approach for accurate fake news detection. Journal of King Saud University – Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2025.xxx

Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. Journal of Computational Science.

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9. https://doi.org/10.1002/spy2.9

Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems, 274, 110642.

Alghamdi, J., et al. (2024). Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. Engineering Applications of Artificial Intelligence.

Alghamdi, J., et al. (2025). ABERT: Adapting BERT model for efficient detection of human and AI-generated fake news. International Journal of Information Management Data Insights.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236.

Arık, A. O., Parlayandemir, G., & Çelik, S. (2026). LLM-based data augmentation for text classification on imbalanced datasets: A case study on fake news detection. Egyptian Informatics Journal, 33, 100886.

Das, S., Kumari, R., & Singh, R. K. (2025). Detection of fake news by convolutional neural networks and recurrent neural networks. Procedia Computer Science (ICMLDE 2023).

Dahou, A., et al. (2024). Enhancing model performance through translation-based data augmentation in the context of fake news detection. Procedia Computer Science, 244, 342–352.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

Embark, O. (2025). Deep learning for fake news detection: Analysing Facebook's misinformation networks. Procedia Computer Science (MobiSPC 2025).

Gemeda Yigezu, M., et al. (2024). Ethio-Fake: Cutting-edge approaches to combat fake news in under-resourced languages using Explainable AI. Proceedings of CL 2024.

Hosea, I. G., et al. (2023). A machine learning approach to fake news detection using SVM and unsupervised learning model. Proceedings of the Cyber Secure Nigeria Conference 2023, 11–18.

Huang, Y., & Liu, B. (2026). Multi-feature graphs and contrastive learning for rumor detection on social media. Machine Learning with Applications, 24, 100889.

Jin, H., & Wang, P. (2025). Fake news detection on social media using triple-attention mechanism optimized by advanced tailor optimization algorithm. Egyptian Informatics Journal, 32, 100815.

Jouhar, J., Pratap, A., Tijo, N., & Mony, M. (2024). Fake news detection using Python and machine learning. Proceedings of ICIDCA 2024.

Kamaruddin, N. K., et al. (2025). Machine learning technique for online fake news detection. Journal of Governance and Integrity, 8(1), 867–873.

Khan, J. Y., et al. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.

Khanday, A. M. U. D., et al. (2020). Detecting Twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. Social Network Analysis and Mining.

Kondamudi, M. R. et al. (2025). An efficient hybrid Hopfield convolutional neural network for detecting spam bots in Twitter platform. Artificial Intelligence in Agriculture (KeAi).
Lazer, D. M., et al. (2018). The science of fake news. Science, 359(6380), 1094–1096.
Li, W. (2025). Fake news detection: A review of machine learning and deep learning methods. Proceedings of CL 2025.
Nair, V., Pareek, J., & Sanskrit, S. (2024). A knowledge-based deep learning approach for automatic fake news detection using BERT on Twitter. Procedia Computer Science (ICMLDE 2023).
Patel, A. (2024). Machine learning-based detection of fake product reviews and news articles. American Academic Scientific Research Journal for Engineering, Technology, and Sciences.
Paulin, M. A. C., & Balaba, E. I. (2025). Distinguishing truth from deception: A machine learning approach to fake news detection. International Journal of Latest Technology in Engineering, Management & Applied Science, XIV(V).
Rao, D., Miao, X., Jiang, Z., & Li, R. (2024). Addressing vaccine misinformation on social media by leveraging transformers and user association dynamics. (ICCSCEI 2024), 235, 1803–1813.
Reddy, J., Mundra, S., Mundra, A., & Kumar, A. (2024). Ensembling deep learning models for fake news classification. Procedia Computer Science (ICMLDE 2023).
Sabapathy, P. A. R. (2025). A review on machine learning techniques for fake news detection. International Journal of Computer Science and Information Security, 23(3).
Satria, H., et al. (2025a). Comparative of convolutional neural network and support vector machine for fake news detection. Proceedings of ICCSCEI 2025.
Satria, H., et al. (2025b). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications.
Suresh, G. V., et al. (2024). Different machine learning approaches for fake news detection. International Journal of Computer Science and Mechatronics, 10(2).
Tanaja, R. N., et al. (2025). Fake news detection using machine learning: Integrating FakeBERT classification, style analysis, and credibility verification. (ICCSCEI 2025).
Thamiliny, S., et al. (2025). Automated fake news detection using machine learning and NLP techniques. Proceedings of Horizon Interdisciplinary Research Symposium 2025.
Twum, S. (2025). Misinformation detection on social media: A big data and machine learning approach. International Journal of Science, Architecture, Technology, and Environment, 2(5).
Vasist, P. N., & Sebastian, M. P. (2022). Tackling the infodemic during a pandemic: A comparative study on algorithms to deal with thematically heterogeneous fake news. International Journal of Information Management Data Insights, 2(2), 100133.
Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.
Wasim, M., et al. (2026). KeepUp: A unified framework fusing knowledge extraction, social platform engagement, and user profiling for fake news detection. Array, 29, 100687.
Wiese, T., & Wiese, J. (2022). Supervised machine learning applications for detecting Russian Federation sponsored social media misinformation. Journal of Big Data.
Yousif, J. H. (2025). Artificial intelligence and machine learning for enhancing Arabic fake news detection: A BERT-based transformer approach. Proceedings of CL 2025.
Zeeshan, H. M., et al. (2025). A machine learning-based scientometric evaluation for fake news detection. ICCK Transactions on Intelligent Systematics, 2(1), 38–48.
Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40.