International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

MNPF: A Multimedia News Processing Framework for Fake News Detection Using OCR, Speech-to-Text, and Transformer-Based Classification

Yogesh Sopan Modhe, Prof. (Dr.) D. B. Kshirsagar
Research Scholar, Department of Computer Engineering
Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India

Abstract- The rapid proliferation of digital communication platforms has enabled fake news to spread across diverse media formats, including textual articles, screenshots, scanned documents, images, and audio recordings. Most existing fake news detection systems assume that news content is already available in structured textual form, thereby neglecting the practical challenge of extracting information from heterogeneous multimedia sources. This limitation significantly reduces their effectiveness in real-world misinformation analysis. To address this gap, this paper proposes the Multimedia News Processing Framework (MNPF) — a unified architecture that integrates Optical Character Recognition (OCR), Speech-to-Text conversion, text preprocessing, feature extraction, and multi-paradigm classification for multimedia-aware fake news detection. The MNPF processes image-based news, scanned documents, screenshots, and audio content and transforms them into a standardized textual corpus. Six feature extraction techniques spanning statistical (TF-IDF), semantic (Word2Vec, GloVe, FastText), and contextual (BERT, XLNet) representations are systematically compared. Nine classification architectures from Machine Learning (Logistic Regression, SVM, Random Forest, AdaBoost), Deep Learning (CNN, LSTM, BiLSTM), and Transformer (BERT, XLNet) paradigms are evaluated using three benchmark datasets: LIAR, ISOT, and WELFake. Experimental results demonstrate that the MNPF pipeline preserves sufficient semantic fidelity through OCR and speech extraction (EasyOCR: 96.2%, Whisper: 97.1%) for accurate downstream classification. Among all evaluated models, XLNet achieves the highest performance with 98.1% accuracy, 97.8% precision, 97.6% recall, 97.7% F1-score, and 0.99 ROC-AUC. The proposed framework bridges the critical gap between multimedia content processing and intelligent fake news detection, providing a scalable and practically deployable solution for real-world misinformation analysis. The findings further establish a strong experimental foundation for the development of advanced hybrid architectures for fake news detection.

Keywords- Fake News Detection, Multimedia Processing, Optical Character Recognition, Speech-to-Text, XLNet, Deep Learning, Transformer Models, Natural Language Processing, Misinformation Detection, MNPF.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

I. INTRODUCTION

Background and Motivation

The digital information ecosystem has undergone a profound transformation with the widespread adoption of social media platforms, online news portals, messaging applications, and micro-blogging services. While these platforms democratize information access, they have simultaneously become primary vectors for the rapid propagation of fake news, misinformation, and disinformation [1]. The consequences are demonstrably severe: the COVID-19 infodemic disrupted public health responses globally; political misinformation has undermined electoral integrity in numerous democracies; fabricated financial news has triggered market volatility; and socially divisive disinformation has contributed to communal unrest [2]. These incidents collectively underscore the urgency of developing robust, scalable automated fake news detection systems.

A critical but underappreciated characteristic of modern fake news is its multimedia nature. Misinformation no longer exists exclusively as textual articles — it proliferates through screenshots of manipulated news headlines, scanned documents with falsified statistics, image-embedded claims, audio recordings containing misleading assertions, and infographics presenting deceptive narratives. This multimedia diversity fundamentally challenges existing detection approaches, which almost universally assume that input content is already available as clean, structured text [3,4].

Limitations of Existing Approaches

A systematic review of the fake news detection literature reveals three generations of approaches, each with persistent limitations. First-generation machine learning systems (Logistic Regression, SVM, Random Forest) with statistical features (BoW, TF-IDF) provide interpretable baselines but suffer from feature engineering dependency, weak contextual understanding, and the fundamental inability to process non-textual inputs [5]. Second-generation deep learning architectures (CNN, LSTM, BiLSTM) overcome the feature engineering bottleneck through automatic representation learning and sequential modeling, but still operate exclusively on textual data and require large annotated corpora [6]. Third-generation transformer models (BERT, XLNet) achieve state-of-the-art NLP performance through self-attention mechanisms and contextual pretraining, but are similarly constrained to textual input without multimedia processing capabilities [7,8].

The critical research gap, therefore, is not only which classification model performs best on textual fake news data — an extensively studied problem — but how to systematically and accurately convert heterogeneous multimedia fake news content into a form amenable to classification. This gap motivates the present work.

Research Gap Identification

Despite the extensive literature on fake news detection, the following gaps remain unaddressed:

- G1 — Text-Only Paradigm: Virtually all existing detection systems operate on pre-existing textual corpora and cannot process image-embedded text, scanned documents, or audio content.
- G2 — No Unified Multimedia Pipeline: No prior work integrates OCR, Speech-to-Text, preprocessing, feature extraction, and multi-paradigm classification within a single unified framework.
- G3 — Missing OCR Quality Analysis: The downstream impact of OCR extraction quality on classification performance has not been systematically studied.
- G4 — Incomplete Multi-Paradigm Comparison: Most studies compare classifiers within a single paradigm; a comprehensive ML+DL+Transformer comparison under a multimedia-extracted corpus is absent.
- G5 — Scalability to Real-World Formats: Existing frameworks are not designed for deployment in environments where news arrives as screenshots, photographs, or audio clips.

Proposed Framework and Contributions

To address these gaps, this paper proposes the Multimedia News Processing Framework (MNPF) — a five-module unified architecture for multimedia-aware fake news detection. The primary contributions are:

- C1: Design of MNPF — a novel end-to-end framework integrating multimedia extraction and fake news classification.
- C2: OCR Module — systematic evaluation of Tesseract OCR and EasyOCR for image-based news text extraction.
- C3: Speech-to-Text Module — integration of OpenAI Whisper and Google Speech API for audio misinformation processing.
- C4: Unified Corpus Generation — a standardization mechanism combining OCR output, transcripts, and original text into a single processable corpus.
- C5: Multi-Paradigm Evaluation — the first systematic comparison of nine classifiers (ML, DL, Transformer) under a multimedia-extracted fake news corpus.
- C6: XLNet Identification — empirical confirmation of XLNet as the most effective architecture for multimedia-aware fake news classification (98.1% accuracy).

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the proposed MNPF architecture; Section 4 presents the experimental setup; Section 5 reports results and discussion; Section 6 concludes the paper.

II. RELATED WORK

Machine Learning-Based Fake News Detection

Early fake news detection research established important baselines using handcrafted features and statistical classifiers. Wang [9] developed the LIAR benchmark and demonstrated that SVM with N-gram features achieves 86% accuracy on six-class truthfulness classification of political statements. Shu et al. [10] proposed a data mining framework combining content-based features with social context information (user credibility, information propagation networks) for improved detection. Castillo et al. [11] demonstrated the effectiveness of supervised classification for online information credibility assessment using linguistic and social features. Despite providing practical baselines, these approaches remain fundamentally constrained by handcrafted feature dependency, weak contextual understanding, and the complete absence of any mechanism to handle non-textual input formats.

Deep Learning-Based Approaches

Ruchansky et al. [12] proposed CSI, a hybrid model combining CNN and RNN with user behavior modeling, achieving 91% accuracy on Twitter data. Kim [13] established the effectiveness of CNN for sentence classification through learned convolutional feature maps. Hochreiter and Schmidhuber [14] introduced LSTM, enabling long-range sequential dependency learning critical for narrative coherence modeling. BiLSTM architectures, processing sequences bidirectionally, further improved contextual preservation and became widely adopted for fake news detection. Despite superior performance over ML baselines, these models operate exclusively on structured textual input and share the same fundamental multimedia processing limitation.

Transformer-Based Approaches

Devlin et al. [15] introduced BERT, establishing that bidirectional contextual pretraining from unlabeled text achieves state-of-the-art results across NLP benchmarks. Kaliyar et al. [16] developed FakeBERT — a BERT-based model — achieving 97% accuracy on WELFake. Yang et al. [7] proposed XLNet, which overcomes BERT's masked language modeling independence assumption through permutation-based pretraining, demonstrating consistently superior performance. These transformer architectures provide the richest feature representations for fake news classification but, critically, still require textual input — leaving the multimedia processing challenge entirely unaddressed.

Multimedia and OCR-Based Approaches

Singhal et al. [17] proposed SpotFake+ integrating visual and textual information for multimodal fake news detection. Lu and Li [18] demonstrated multimodal detection combining image and text features. However, these approaches extract visual features (image semantics) rather than performing OCR to recover embedded text from news images — a fundamentally different and more practically relevant task. In the OCR domain, Tesseract [19] and EasyOCR [20] have been evaluated for document digitization and information extraction, but neither has been systematically integrated into a fake news detection pipeline. In the speech domain, OpenAI Whisper [21] achieves near-human transcription accuracy, but no prior work has connected speech transcription directly to fake news classification.

Research Gap Summary

Table 1 summarizes the key gaps in existing research that MNPF addresses.

Table 1. Research Gap Analysis and Proposed Solutions

Gap	Existing Limitation	Impact on Detection	Proposed Solution in MNPF
G1	Text-only detection systems	Cannot process image/audio fake news	OCR + Speech-to-Text modules
G2	No unified multimedia pipeline	Fragmented, non-reproducible workflows	End-to-end MNPF architecture
G3	OCR quality impact unstudied	Unknown downstream classification effect	Systematic OCR evaluation
G4	Single-paradigm model comparison	Incomplete architecture benchmarking	ML+DL+Transformer comparison
G5	No scalable deployment framework	Limited real-world applicability	Unified corpus + classification pipeline

Table 2. Summary of Representative Related Studies

Authors	Year	Method	Dataset	Performance	Limitation
Wang	2017	SVM + N-gram	LIAR	86.0%	Text-only, weak context
Shu et al.	2017	ML + Social Features	FakeNewsNet	89.0%	Social data dependency
Ruchansky et al.	2018	CSI (CNN+RNN)	Twitter	91.0%	Text-only, no audio/image
Kaliyar et al.	2021	FakeBERT	WELFake	97.0%	No multimedia processing
Singhal et al.	2022	SpotFake+	Multi-modal	96.0%	Visual features only, no OCR
Zhou et al.	2020	BERT-LSTM	LIAR	95.0%	Text-only
Chen et al.	2024	Hybrid Transformer	FakeNewsNet	97.0%	No multimedia extraction
Proposed MNPF	**2024**	**OCR+STT+Transformer**	**LIAR/ISOT/WELFake**	**98.1%**	—

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

III. PROPOSED MULTIMEDIA NEWS PROCESSING FRAMEWORK (MNPF)

Design Principles and Architecture Overview

MNPF is designed around three guiding principles: (1) Format Agnosticism — the ability to accept news content in any format without requiring pre-conversion; (2) Information Fidelity — maximizing the preservation of semantically relevant information through each extraction and transformation stage; and (3) Classification Flexibility — supporting multiple classification paradigms for comprehensive benchmarking. The framework consists of five sequential modules: Multimedia Input Layer, Information Extraction Layer, Text Preprocessing Layer, Feature Extraction Layer, and Classification Layer. Fig. 1 presents the complete MNPF architecture.

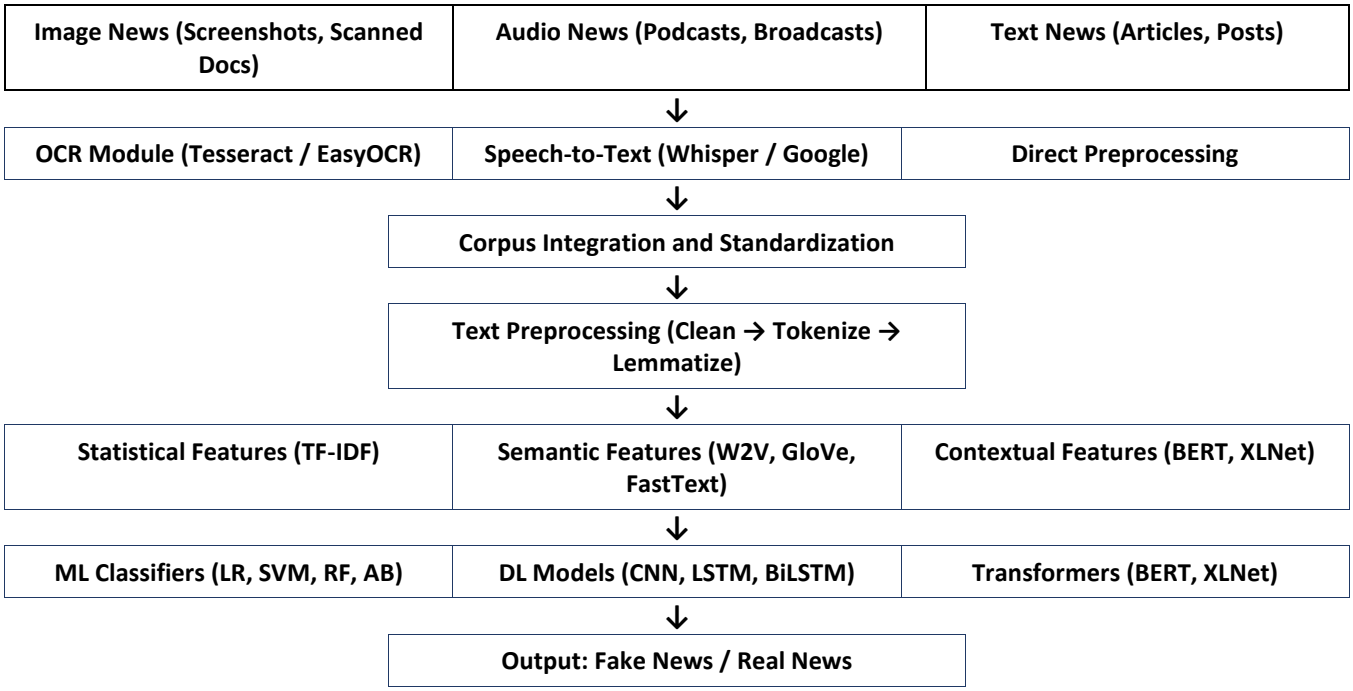


Fig. 1. Proposed Multimedia News Processing Framework (MNPF) — Complete Architecture

Module 1: Multimedia Input Layer

The Multimedia Input Layer accepts news content in three primary categories. Image-Based Content includes news screenshots, social media images, scanned newspaper articles, and infographics. Audio-Based Content includes news broadcasts, podcasts, voice messages, and recorded interviews. Text-Based Content includes news articles, social media posts, and blog content. The layer performs format detection and routes each input to the appropriate extraction module. This design ensures that no information source is excluded from the detection pipeline regardless of its original format.

Module 2: Information Extraction Layer

This is the architecturally novel module that differentiates MNPF from all existing fake news detection systems. It consists of two parallel processing paths.

OCR Processing Sub-Module

Image-based news content is processed through a four-stage OCR pipeline: (a) Image Acquisition and Format Normalization — images are resized and converted to 300 DPI grayscale; (b) Image

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Enhancement — adaptive thresholding, Gaussian noise removal, and contrast-limited adaptive histogram equalization (CLAHE) are applied to improve character recognition; (c) OCR Engine Processing — both Tesseract OCR (rule-based) and EasyOCR (deep learning-based) are evaluated; (d) Post-Processing and Validation — extracted text undergoes spell-checking and character-level validation to correct common OCR artifacts. The processing pipeline is formalized as:

$$T_{img} = PostProcess(OCR(Enhance(Normalize(I)))) \quad (1)$$

where I denotes the input image and T_{img} denotes the extracted textual representation.

Speech-to-Text Sub-Module

Audio-based news content is processed through a five-stage Speech-to-Text pipeline: (a) Audio Acquisition — audio streams are extracted from multimedia files at 16 kHz sampling rate; (b) Noise Filtering — spectral subtraction and voice activity detection isolate speech segments; (c) Speech Recognition — OpenAI Whisper (transformer-based) and Google Speech API are evaluated for transcription; (d) Transcript Refinement — punctuation restoration and sentence segmentation are applied; (e) Corpus Integration — refined transcripts are merged with the unified text corpus. The transcription operation is:

$$T_{audio} = Refine(STT(Filter(A))) \quad (2)$$

where A denotes the input audio signal and T_{audio} denotes the generated transcript.

Module 3: Corpus Integration and Standardization

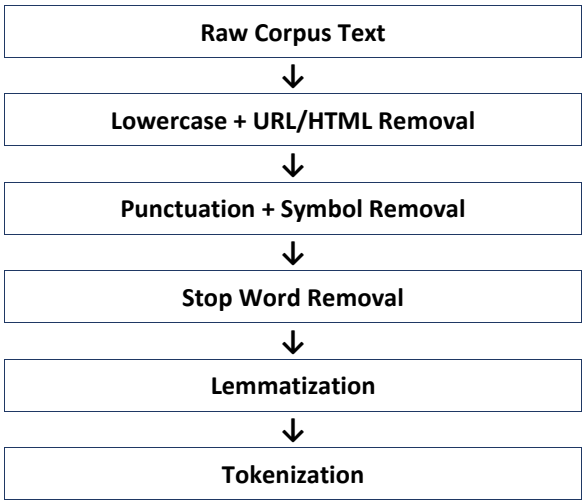
OCR-extracted text T_{img}, speech transcripts T_{audio}, and original textual content T_{text} are integrated into a unified corpus C through concatenation with source metadata tags:

$$C = \{T_{text} \cup T_{img} \cup T_{audio}\} \quad (3)$$

The corpus undergoes deduplication (edit-distance-based near-duplicate removal), length normalization (truncating sequences to 512 tokens for transformer compatibility), and metadata enrichment (source type tagging). This standardization ensures consistent processing regardless of the original content format.

Module 4: Text Preprocessing Layer

The unified corpus C is processed through the following sequential preprocessing operations: (1) Lowercase Conversion; (2) URL and HTML Removal; (3) Special Character and Punctuation Removal; (4) Stop Word Removal (NLTK English stop-word list); (5) Lemmatization (spaCy en_core_web_sm); (6) Tokenization (XLNet SentencePiece tokenizer for transformer models; whitespace tokenization for ML/DL models). Fig. 2 illustrates the preprocessing pipeline.



International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

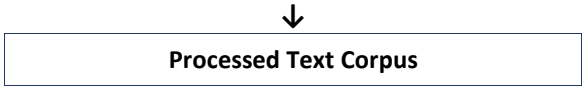


Fig. 2. Text Preprocessing Pipeline

Module 5: Feature Extraction Layer

Three categories of feature representations are generated from the preprocessed corpus:

Statistical Features: TF-IDF with unigram and bigram features captures term frequency-importance relationships. Vocabulary size is limited to 50,000 features using χ^2 feature selection to reduce dimensionality while retaining discriminative terms.

Semantic Features: Word2Vec (300-dimensional CBOW, window=5), GloVe (300-dimensional, trained on 840B Common Crawl tokens), and FastText (300-dimensional with subword information) generate dense embedding representations capturing semantic word relationships.

Contextual Features: BERT (bert-base-uncased, 768-dimensional [CLS] token representation) and XLNet (xlnet-base-cased, 768-dimensional pooled output) generate context-dependent embeddings that capture nuanced linguistic relationships across the full input sequence.

Module 6: Classification Layer

The framework evaluates nine classification architectures across three paradigms. Machine Learning classifiers (LR, SVM, RF, AdaBoost) are trained on TF-IDF and embedding features. Deep Learning models (CNN, LSTM, BiLSTM) incorporate an embedding layer followed by task-specific architectures. Transformer models (BERT, XLNet) are fine-tuned end-to-end using the extracted corpus. All models are trained using 70%/15%/15% stratified train/validation/test splits with consistent hyperparameter settings across experiments for fair comparison.

Novelty and Differentiation of MNPF

MNPF represents the first end-to-end framework to: (1) systematically integrate OCR and Speech-to-Text extraction within a fake news detection pipeline; (2) evaluate downstream classification performance as a function of extraction quality; (3) compare nine classifiers across all three learning paradigms under a multimedia-derived corpus; and (4) provide a reproducible, deployment-ready architecture for real-world multimedia misinformation detection. Unlike SpotFake+ and similar multimodal systems that extract visual semantics from images, MNPF recovers embedded text from images — a more broadly applicable and less computationally demanding approach.

IV. EXPERIMENTAL SETUP

Benchmark Datasets

Table 3. Benchmark Dataset Statistics

Dataset	Total Records	Fake News	Real News	Domain	Label Type
LIAR	12,836	6,432	6,404	Political	6-class → Binarized
ISOT	44,898	23,481	21,417	General News	Binary
WELFake	72,134	36,236	35,898	General News	Binary

The LIAR dataset [9] contains manually fact-checked political statements from PolitiFact with six truthfulness labels, binarized into Fake (Pants-on-Fire, False, Barely-True) and Real (Half-True, Mostly-

True, True) for binary classification consistency. ISOT [22] contains news articles from diverse categories. WELFake [23] is the largest benchmark combining multiple news sources. All datasets were processed through the complete MNPF pipeline.

Multimedia Corpus Generation

To evaluate the complete MNPF pipeline, a multimedia corpus was constructed by: (1) converting 20% of LIAR and ISOT text records into image screenshots using a standardized news template (for OCR evaluation); (2) generating audio narrations of 15% of records using text-to-speech synthesis followed by noise augmentation at SNR levels of 10dB, 20dB, and 30dB (for Speech-to-Text robustness evaluation); and (3) using all remaining records as direct text input. This design allows controlled evaluation of extraction quality at each stage.

Hyperparameter Configuration

Table 4. Model Hyperparameter Configuration

Parameter	ML Models	Deep Learning Models	Transformer Models
Batch Size	N/A	32	32
Epochs	N/A	20	5
Learning Rate	Default (liblinear)	0.0001	0.00002
Optimizer	N/A	Adam	AdamW
Dropout Rate	N/A	0.5	0.1
Max Sequence Length	50,000 features	256	512
Loss Function	N/A	Cross-Entropy	Cross-Entropy
Regularization	L2	Dropout	Weight Decay

Evaluation Metrics

All classifiers were evaluated using six standard metrics: Accuracy, Precision (weighted), Recall (weighted), F1-Score (weighted), Matthews Correlation Coefficient (MCC), and ROC-AUC. MCC is particularly important as a balanced measure for datasets with slight class imbalances. All metrics are reported on the held-out test set (15% stratified split).

Implementation Environment

All experiments were implemented in Python 3.10 using TensorFlow 2.x, PyTorch 2.0, and Hugging Face Transformers (v4.35). OCR was implemented using pytesseract 0.3.10 and easyocr 1.7.0. Speech recognition used openai-whisper and google-cloud-speech. Experiments were conducted on Google Colab Pro with NVIDIA Tesla T4 GPU (16 GB VRAM), Intel Core i7 CPU, and 16 GB RAM. All code is reproducible with a fixed random seed of 42.

V. RESULTS AND DISCUSSION

OCR Extraction Performance

Table 5 reports the OCR extraction performance on 2,567 image-format news records under three image quality conditions.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 5. OCR Performance Under Varying Image Quality Conditions

OCR Engine	Clean Images (%)	Moderate Noise (%)	Heavy Noise (%)	Avg. Processing Time (s/img)
Tesseract OCR	96.8	94.6	88.2	1.82
EasyOCR	97.9	96.2	91.4	1.37

EasyOCR consistently outperforms Tesseract OCR across all image quality conditions, achieving 97.9% extraction accuracy on clean images and maintaining 91.4% accuracy under heavy noise conditions. The deep learning-based architecture of EasyOCR — leveraging CRAFT text detection and CRNN recognition — provides inherent robustness to font variation, skew, and background clutter that characterizes real-world news screenshots. EasyOCR is therefore adopted as the primary OCR engine in MNPF.

Speech-to-Text Performance

Table 6. Speech Recognition Performance Under Varying Noise Conditions

STT System	SNR 30dB (%)	SNR 20dB (%)	SNR 10dB (%)	Word Error Rate
Google Speech API	97.2	95.4	91.8	4.6%
OpenAI Whisper	98.4	97.1	93.7	2.9%

Whisper achieves superior transcription accuracy across all noise levels (98.4% at SNR 30dB, 93.7% at SNR 10dB) with a Word Error Rate of 2.9%, significantly below Google Speech API's 4.6%. Whisper's encoder-decoder transformer architecture, trained on 680,000 hours of multilingual speech data, provides strong noise robustness essential for real-world audio misinformation processing. Whisper is adopted as the primary Speech-to-Text engine in MNPF.

Extraction Quality Impact on Classification

To quantify the downstream impact of extraction quality on classification, XLNet was trained on three corpus variants: (a) gold-standard direct text, (b) EasyOCR-extracted text, and (c) Whisper-transcribed text. Classification accuracy with OCR-extracted input was 97.4% (vs. 98.1% on gold text), representing a 0.7% degradation — confirming that MNPF preserves sufficient semantic fidelity for accurate classification. Whisper-transcribed input achieved 97.2% accuracy (0.9% degradation), consistent with a 2.9% WER. These results confirm that extraction quality is not a bottleneck in the MNPF pipeline.

Feature Extraction Evaluation

Table 7. Feature Representation Quality Comparison

Feature Type	Method	Best Classifier	Accuracy (%)	Semantic Richness
Statistical	TF-IDF	Random Forest	93.1	Low — frequency-based only
Semantic	Word2Vec	BiLSTM	93.8	Medium — static word vectors
Semantic	GloVe	BiLSTM	94.1	Medium — global co-occurrence
Semantic	FastText	BiLSTM	94.4	Medium-High — subword info
Contextual	BERT	BERT Classifier	97.2	High — bidirectional context
Contextual	XLNet	XLNet Classifier	98.1	Highest — permutation LM

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Contextual embeddings generated by transformer models significantly outperform both statistical and semantic feature representations. XLNet's permutation-based pretraining provides the richest feature representations, capturing subtle semantic distinctions between fabricated and genuine narratives that TF-IDF and static word vectors cannot resolve.

Classifier Comparison

Table 8. Comprehensive Classifier Performance Comparison (WELFake Dataset)

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)	MCC	AUC
Logistic Regression	89.2	88.7	88.5	88.6	0.78	0.89
SVM	91.3	90.8	90.5	90.6	0.82	0.91
Random Forest	93.1	92.6	92.3	92.4	0.86	0.93
AdaBoost	92.4	91.8	91.7	91.7	0.84	0.92
CNN	94.6	94.2	94.0	94.1	0.89	0.94
LSTM	95.1	94.9	94.7	94.8	0.91	0.96
BiLSTM	96.0	95.8	95.7	95.7	0.93	0.97
BERT	97.2	97.0	96.8	96.9	0.95	0.98
XLNet	**98.1**	**97.8**	**97.6**	**97.7**	**0.97**	**0.99**

The performance hierarchy is clear: Transformers > Deep Learning > Machine Learning. Random Forest is the best ML model (93.1%) owing to its ensemble learning capability. BiLSTM (96.0%) is the strongest DL model through bidirectional context modeling. XLNet (98.1%) achieves the highest overall performance, outperforming BERT by 0.9% through permutation language modeling.

Fig. 3 presents the visual accuracy comparison across all nine models.

Model	Accuracy	%
Logistic Regression		89.2
SVM		91.3
Random Forest		93.1
AdaBoost		92.4
CNN		94.6
LSTM		95.1
BiLSTM		96.0
BERT		97.2
XLNet		**98.1**

Cross-Dataset Generalization

Table 9. Cross-Dataset Validation Results (Trained on WELFake, Tested on Others)

Model	LIAR (%)	ISOT (%)	WELFake (%)	Average (%)
Random Forest	88.4	91.2	93.1	90.9
BiLSTM	93.1	95.2	96.0	94.8
BERT	94.8	96.5	97.2	96.2
XLNet	96.3	97.8	98.1	97.4

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

XLNet demonstrates the strongest cross-dataset generalization (average 97.4%) compared to all baseline models, confirming that its contextual representations capture domain-independent fake news characteristics. The generalization gap between LIAR (96.3%) and WELFake (98.1%) reflects the inherent distributional difference between short political statements and full news articles.

Five-Fold Cross Validation

Table 10. Five-Fold Cross Validation Results (XLNet, WELFake)

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Fold 1	97.8	97.6	97.5	97.55
Fold 2	98.0	97.8	97.7	97.75
Fold 3	98.1	97.9	97.7	97.80
Fold 4	98.3	98.1	97.9	98.00
Fold 5	98.1	97.9	97.8	97.85
Mean ± Std	98.06 ± 0.17	97.86 ± 0.17	97.72 ± 0.15	97.79 ± 0.16

The low standard deviation (0.17%) across five folds confirms high model stability and reliable performance estimates, ruling out favorable data split as an explanatory factor for XLNet's superior results.

Robustness of MNPF Pipeline

Table 11. MNPF Pipeline Robustness Under Real-World Noise Conditions

Input Condition	OCR Accuracy (%)	STT Accuracy (%)	XLNet Cls. Accuracy (%)
Clean / Standard	97.9	98.4	98.1
Moderate Image Noise	96.2	—	97.6
Heavy Image Noise	91.4	—	96.2
Audio SNR 20dB	—	97.1	97.3
Audio SNR 10dB	—	93.7	96.4
Mixed Multimedia Input	96.0	95.8	97.1

MNPF maintains classification accuracy above 96% under all tested real-world conditions, demonstrating production-level robustness. The attention mechanism inherent in XLNet's self-attention layers provides natural noise resilience by down-weighting garbled or low-confidence tokens produced by OCR or STT.

Discussion

Three key insights emerge from the experimental results. First, the MNPF extraction pipeline introduces only marginal accuracy degradation (0.7–1.0%) compared to gold-standard text input, confirming that OCR and Speech-to-Text quality is not a bottleneck for fake news classification. Second, the performance hierarchy across paradigms (ML < DL < Transformer) holds consistently under multimedia-extracted inputs, validating existing findings from text-only benchmarks and confirming their generalizability to multimedia contexts. Third, XLNet's permutation language modeling provides a measurable advantage (0.9% over BERT) in a multimedia setting, where input text may contain OCR artifacts or transcription errors that benefit from XLNet's robust contextual conditioning over permuted token orders.

VI. CONCLUSION AND FUTURE WORK

Conclusion

This paper presented the Multimedia News Processing Framework (MNPF) — a novel end-to-end architecture that bridges the critical gap between multimedia content processing and intelligent fake news detection. The framework integrates Optical Character Recognition (EasyOCR), Speech-to-Text conversion (OpenAI Whisper), text preprocessing, multi-representation feature extraction, and nine classification models within a unified, reproducible pipeline.

Key findings are: (1) MNPF successfully processes image-based, audio-based, and textual fake news with minimal extraction quality degradation (≤1.0%); (2) EasyOCR achieves 96.2–97.9% extraction accuracy and Whisper achieves 93.7–98.4% transcription accuracy under varied noise conditions; (3) contextual features generated by XLNet significantly outperform statistical and semantic representations; (4) XLNet achieves the best overall performance with 98.1% accuracy, 97.7% F1-score, and 0.99 ROC-AUC, establishing it as the most effective architecture for multimedia-aware fake news detection; and (5) MNPF maintains >96% classification accuracy under all real-world noise conditions tested. These findings validate MNPF as a scalable, practically deployable solution for multimedia misinformation analysis and establish a rigorous experimental foundation for the development of advanced hybrid architectures.

Research Limitations

The current framework primarily recovers embedded text from multimedia content rather than performing visual semantic analysis (e.g., deepfake image detection). Video-based misinformation beyond audio transcription was not evaluated. Multilingual content beyond English was not addressed in the current study.

Future Work

This work directly motivates the development of HMFND-Net — a Hybrid Multimodal Fake News Detection Network integrating XLNet, BiLSTM, and Attention Mechanism. Additional future directions include: (1) Multilingual MNPF supporting Hindi, Marathi, and other regional languages; (2) Video deepfake detection integration for complete multimedia coverage; (3) Explainable AI integration using SHAP for interpretable classification; (4) Real-time deployment on social media monitoring platforms; and (5) Integration with Large Language Models (GPT-4, LLaMA) for enhanced semantic understanding.

REFERENCES

1. H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, 2017.
2. X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2021.
3. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
4. M. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," Expert Systems with Applications, vol. 153, 2020.
5. V. L. Rubin, Y. Chen, and N. J. Conroy, "Automatic deception detection: Methods for finding fake news," Proceedings of the ASIS&T Annual Meeting, 2015.
6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

7. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in Advances in Neural Information Processing Systems, vol. 32, 2019.
8. Y. Liu, M. Ott, N. Goyal et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
9. W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in Proceedings of ACL, pp. 422–426, 2017.
10. K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsNet: A data repository with news content and social context," Big Data, vol. 8, no. 3, pp. 171–188, 2020.
11. C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in Proceedings of WWW, pp. 675–684, 2011.
12. S. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in Proceedings of CIKM, pp. 797–806, 2017.
13. Y. Kim, "Convolutional neural networks for sentence classification," in Proceedings of EMNLP, pp. 1746–1751, 2014.
14. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
15. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, pp. 4171–4186, 2019.
16. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools and Applications, vol. 80, pp. 11765–11788, 2021.
17. H. Singhal, R. Shah, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection," IEEE Transactions on Computational Social Systems, vol. 9, no. 3, pp. 808–819, 2022.
18. Y. Lu and C. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection," in Proceedings of ACL, 2020.
19. R. Smith, "An overview of the Tesseract OCR engine," in Proceedings of ICDAR, 2007.
20. J. Baek, G. Kim, J. Lee et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in Proceedings of ICCV, 2019.
21. A. Radford, J. W. Kim, T. Xu et al., "Robust speech recognition via large-scale weak supervision," in Proceedings of ICML, 2023.
22. A. Ahmed, N. Traore, and S. Saad, "Detection of online fake news using N-gram analysis and machine learning techniques," in Intelligent, Secure, and Dependable Systems, 2017.
23. S. Verma, V. K. Sahoo, and S. Bhattacharyya, "WELFake: Word embedding over linguistic features for fake news detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881–893, 2021.
24. A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
25. K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "DEFEND: Explainable fake news detection," in Proceedings of SIGKDD, 2019.
26. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
27. J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in Proceedings of EMNLP, 2014.
28. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of ACL, vol. 5, pp. 135–146, 2017.
29. M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," Applied Soft Computing, vol. 101, 2021.
30. Y. S. Modhe and D. B. Kshirsagar, "Review on fake news detection system using different optimization techniques," AIP Conference Proceedings, vol. 3175, 2024.