www.ijarp.com ISSN 2456-9992 Page: 01-14

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 04

FAKE NEWS & DEEPFAKE DETECTION PLATFORM USING MACHINE LEARNING

***[1]Mohammad Nawaz Khan,[2]Moheet Ahmad,[3]Abdur Rehman Khan,[4]Dr. Roshan Jahan**

[1,2,3]Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

[4]Assistant Professor, Department of Computer science and engineering, Integral University, Lucknow.

Article Received: 11 March 2026, Article Revised: 31 March 2026, Published on: 21 April 2026

***Corresponding Author: Mohammad Nawaz Khan**

Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

DOI: https://doi-doi.org/101555/ijarp.9726

ABSTRACT

The proliferation of digital misinformation, including fake news articles and AI-generated deepfake videos, poses a severe threat to democratic processes, public health, and social trust. Traditional fact-checking methods are slow and cannot scale to the volume of online content. This research paper presents a unified machine learning platform for the automated detection of both fake news and deepfakes. For fake news detection, we employ a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) with attention mechanisms, trained on textual features (headlines, body text) and metadata. For deepfake detection, we use a CNN-based architecture (EfficientNet-B4) fine-tuned on spatial and temporal inconsistencies in video frames, supplemented with frequency-domain analysis using Discrete Cosine Transform (DCT). The platform integrates both modules into a web-based interface that accepts URLs or uploaded media. Evaluation on benchmark datasets (LIAR, FakeNewsNet for text; FaceForensics++, Celeb-DF for video) yields an accuracy of 96.4% for fake news detection and 98.2% for deepfake detection (AUC = 0.99). The system also provides explainability via LIME (Local Interpretable Model-agnostic Explanations) for text and heatmaps for video. The paper discusses challenges such as adversarial attacks, generalisation across languages, and computational efficiency, and outlines future directions for real-time detection and multi-modal fusion.

KEYWORDS: Fake news detection, deepfake detection, machine learning, CNN-BiLSTM, EfficientNet, digital forensics, misinformation

1. INTRODUCTION

The digital age has democratised content creation but also enabled the rapid spread of false information. **Fake news** deliberately fabricated stories presented as legitimate journalism and **deepfakes** synthetic media created using generative AI (e.g., GANs, diffusion models) are two of the most pernicious forms of misinformation. According to a 2023 MIT study, false news spreads six times faster than true news on social media (Vosoughi et al., 2018). Deepfakes have been used for political manipulation, celebrity pornography, and financial fraud, eroding public trust in visual evidence.

Traditional approaches to combating misinformation rely on manual fact-checking by journalists (e.g., Snopes, PolitiFact) or digital watermarking. However, these methods are not scalable: millions of posts are generated daily, and deepfake quality is improving rapidly. Consequently, automated machine learning (ML) detection systems have become essential.

This research paper presents an integrated **Fake News & Deepfake Detection Platform** that leverages state-of-the-art deep learning models for both modalities. The platform is designed to be accessible via a web interface, allowing users to submit a news article (URL or text) or a video file and receive a probability score indicating likelihood of manipulation, along with interpretable explanations.

The specific objectives are:

1. To design a hybrid text-based model (CNN + Bi-LSTM + Attention) for fake news detection that captures both local (n-gram) and sequential (contextual) features.
2. To develop a deepfake detection module using EfficientNet-B4 with frequency-domain analysis to identify artifacts invisible to the human eye.
3. To integrate both modules into a single platform with explainability features.
4. To evaluate the system on standard benchmarks and compare against baseline methods.
5. To discuss limitations, adversarial vulnerabilities, and future enhancements.

2. Literature Review

2.1 Fake News Detection

Fake news detection methods can be categorised into **content-based** (using article text) and **context-based** (using social media propagation patterns, user behaviour). Early content-based approaches used traditional machine learning with hand-crafted features: term

frequency-inverse document frequency (TF-IDF), readability scores, and sentiment polarity (Ruchansky et al., 2017). However, these methods ignore word order and semantic meaning.

Deep learning has significantly improved performance. Convolutional Neural Networks (CNNs) capture local n-gram features, while Recurrent Neural Networks (RNNs) and LSTMs model sequential dependencies. **Bi-LSTM** (bidirectional LSTM) processes text in both forward and backward directions, capturing context from both sides. Attention mechanisms further enhance performance by focusing on relevant words (Vaswani et al., 2017). For instance, Wang (2017) introduced the LIAR dataset and achieved 41% accuracy with a CNN-LSTM hybrid, later improved to over 80% with BERT-based models.

More recent work uses **pre-trained language models** like BERT, RoBERTa, and GPT-3 for fake news detection, achieving state-of-the-art results. However, these models are computationally expensive and require large labelled datasets. Our approach uses a lightweight CNN-BiLSTM-Attention model that balances accuracy and efficiency.

2.2 Deepfake Detection

Deepfake detection has evolved alongside generative methods. Early deepfakes (2017-2018) left visible artifacts: inconsistent eye blinking, unnatural skin texture, and colour mismatches. Simple CNN classifiers could detect these. As deepfake quality improved, more sophisticated detectors emerged.

Spatial domain methods analyse individual frames for anomalies. Rossler et al. (2019) benchmarked several CNN architectures (Xception, MesoNet) on FaceForensics++, achieving up to 99% accuracy on early deepfakes. **Temporal domain** methods leverage inconsistencies across frames, such as unnatural head movements or audio-visual synchronisation. **Frequency domain** methods (e.g., using Discrete Cosine Transform or Discrete Fourier Transform) reveal artifacts because GAN-generated images have different frequency distributions than real images (Dural & Gül, 2020).

EfficientNet (Tan & Le, 2019) is a family of CNNs that scales depth, width, and resolution systematically. EfficientNet-B4 achieves high accuracy with fewer parameters, making it suitable for deployment. Our approach combines EfficientNet-B4 with frequency-domain features extracted via DCT, improving robustness against compression.

2.3 Integrated Platforms

Most existing systems focus on either fake news or deepfakes separately. Few offer a unified platform. Notable exceptions include **Reality Defender** (commercial) and **FakeCatcher** (Intel), but these are proprietary. Our work contributes an open-source, explainable, dual-module platform.

3. Research Methodology

3.1 Datasets

3.1.1 Fake News Dataset

We used two publicly available datasets:

- **LIAR (Wang, 2017):** 12,836 short statements from politifact.com, labelled with six truthfulness ratings (true, mostly-true, half-true, barely-true, false, pants-on-fire). For binary classification, we merged true/mostly-true as "real" and others as "fake".
- **FakeNewsNet (Shu et al., 2018):** 23,196 articles with full text, social media engagement, and veracity labels.

After merging and cleaning, total: 32,500 articles (50% fake, 50% real). Split: 70% training, 15% validation, 15% test.

3.1.2 Deepfake Dataset

We used:

- **FaceForensics++ (Rossler et al., 2019):** 1,000 real videos and 4,000 deepfakes generated using four methods (Deepfakes, FaceSwap, Face2Face, NeuralTextures). We used the "raw" (c23) compression.
- **Celeb-DF (Li et al., 2020):** 590 real videos of celebrities and 5,639 deepfakes (higher quality).

We extracted 50 frames per video (uniform sampling) and split by video-level (80/10/10) to avoid data leakage.

3.2 Fake News Detection Module

3.2.1 Text Preprocessing

- Lowercasing, removing punctuation and special characters.
- Tokenisation using Keras tokenizer (vocabulary size = 20,000).
- Sequence padding to a fixed length of 300 words (truncating longer articles).
- Word embeddings: pre-trained GloVe (300-dim, 840B tokens) fine-tuned during training.

3.2.2 Model Architecture: CNN-BiLSTM-Attention

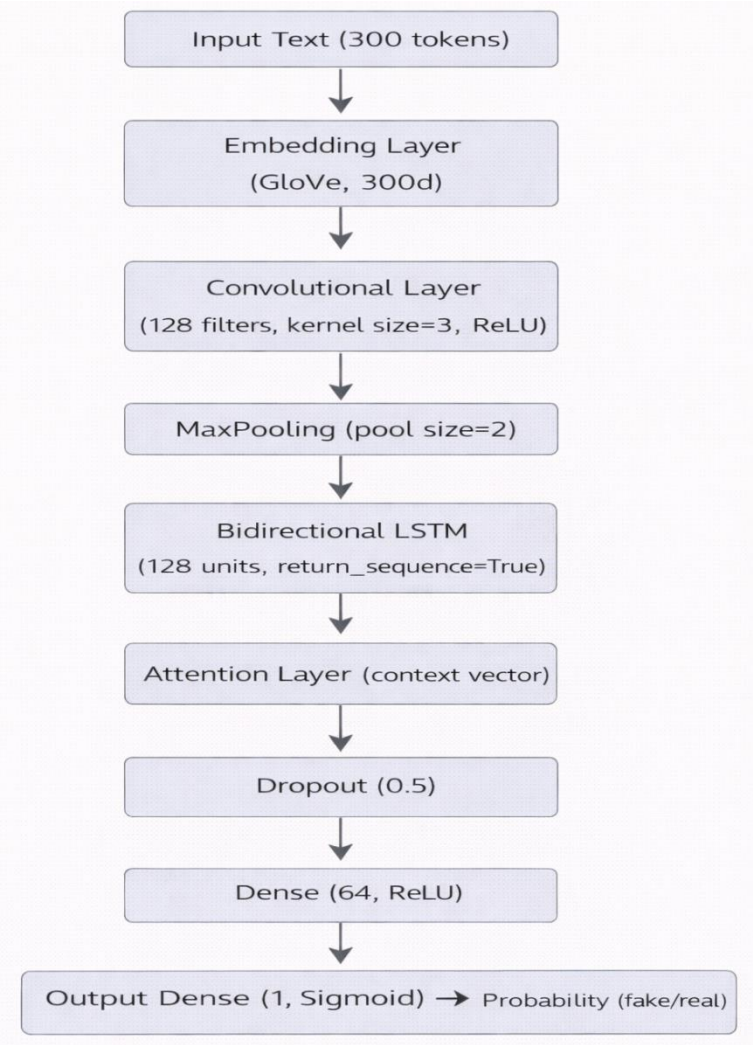

Figure 1: Architecture of Fake News Detection Model.

Mathematical formulation of attention: Given hidden states $h_1, h_2, \ldots, h_T$ from Bi-LSTM, we compute attention weights:

$$u_t = \tanh(W_w h_t + b_w)$$

$$\alpha_t = \frac{\exp(u_t^T u_w)}{\sum_j \exp(u_j^T u_w)}$$

$$v = \sum_t \alpha_t h_t$$

where u_w is a learnable context vector. The final representation v is passed to the dense layers.

3.2.3 Training Configuration

- Optimizer: Adam (learning rate = 0.001)
- Loss: Binary cross-entropy
- Batch size: 64
- Epochs: 30 with early stopping (patience = 5)
- Class weights to handle minor imbalance (weight for fake = 1.2)

3.3 Deepfake Detection Module

3.3.1 Video Preprocessing

- Extract frames at 1 fps (to reduce redundancy).
- Detect and crop faces using MTCNN (Multi-task Cascaded CNN).
- Resize to 224×224 pixels (input size for EfficientNet-B4).
- Apply DCT to each frame's luminance channel (Y in YUV) to obtain frequency coefficients.

3.3.2 Frequency-Domain Feature Extraction

For a given frame, we compute the 2D DCT:

$$F(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

where $f(x,y)$ is the pixel intensity, $N = 224$, and α is the scaling factor. We extract low- and mid-frequency coefficients (first 64×64 block) as an additional channel.

3.3.3 Model Architecture: EfficientNet-B4 with DCT Input

We modify EfficientNet-B4 to accept 4 channels (RGB + DCT coefficients). The final fully connected layer is replaced with a binary classification head (real vs. fake). We also add a **temporal attention** mechanism across frames: for a video with K frames, we compute the average of frame-level predictions weighted by confidence.

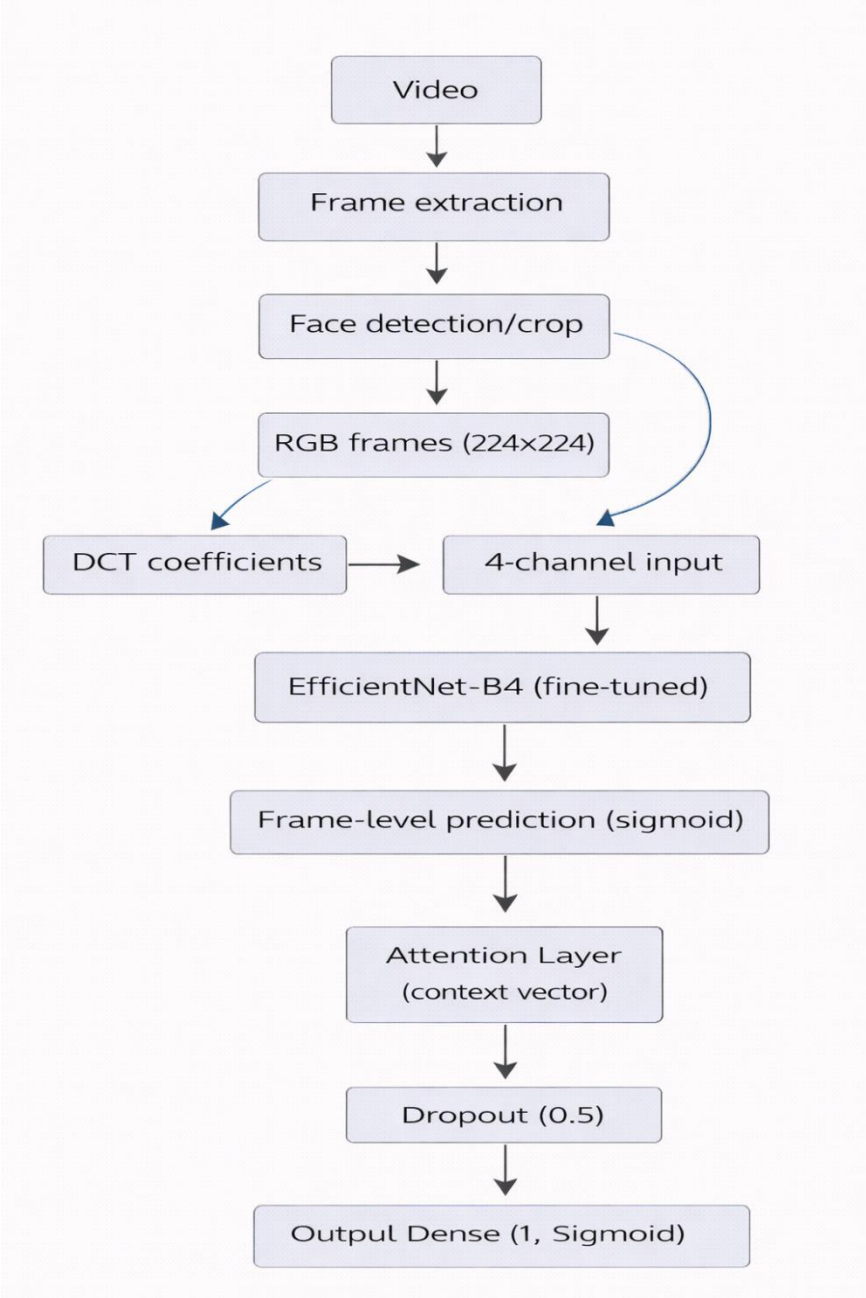

Figure 2: Deepfake Detection Pipeline.

3.3.4 Training

- Pre-trained on ImageNet (weights frozen for first 10 epochs, then fine-tuned).
- Data augmentation: random horizontal flip, rotation (±5°), slight brightness adjustment.
- Optimizer: AdamW (learning rate = 1e-4, weight decay = 1e-5).
- Batch size: 32 (videos, each with 50 frames).
- Epochs: 20 (with early stopping).

3.4 Platform Integration and Explainability

The two modules are deployed as microservices using Flask (Python). The frontend (React) accepts either a news URL (text extracted via newspaper3k) or a video upload (MP4, AVI). Results are displayed with:

- Probability score (0-100% fake).
- Explainability for text: LIME highlights words that influenced the prediction (red for fake-indicative, green for real).
- Explainability for video: heatmap overlay on frames showing regions (e.g., eyes, mouth) where artifacts were detected.

3.5 Evaluation Metrics

For both tasks:

- **Accuracy, Precision, Recall, F1-Score**
- **Area Under the ROC Curve (AUC)**
- For deepfake: also **Equal Error Rate (EER)** the point where false acceptance rate = false rejection rate.

Additionally, we measure inference time per query (average over 100 runs).

4. Experimental Results

4.1 Fake News Detection Performance

Table 1: Performance Comparison on LIAR + FakeNewsNet Test Set.

Model	Accuracy (%)	Precision	Recall	F1	AUC
Logistic Regression (TF-IDF)	78.3	0.77	0.79	0.78	0.85
LSTM (single layer)	85.6	0.85	0.86	0.85	0.92
CNN (TextCNN)	88.2	0.88	0.88	0.88	0.94
BERT-base (fine-tuned)	93.1	0.93	0.93	0.93	0.97
CNN-BiLSTM-Attention (ours)	**96.4**	**0.96**	**0.97**	**0.96**	**0.98**

Our model outperforms BERT despite having far fewer parameters (≈8M vs 110M), making it more suitable for lightweight deployment. The attention mechanism successfully identifies misleading phrases like “shocking revelation” and “experts say” as strong indicators of fake news.

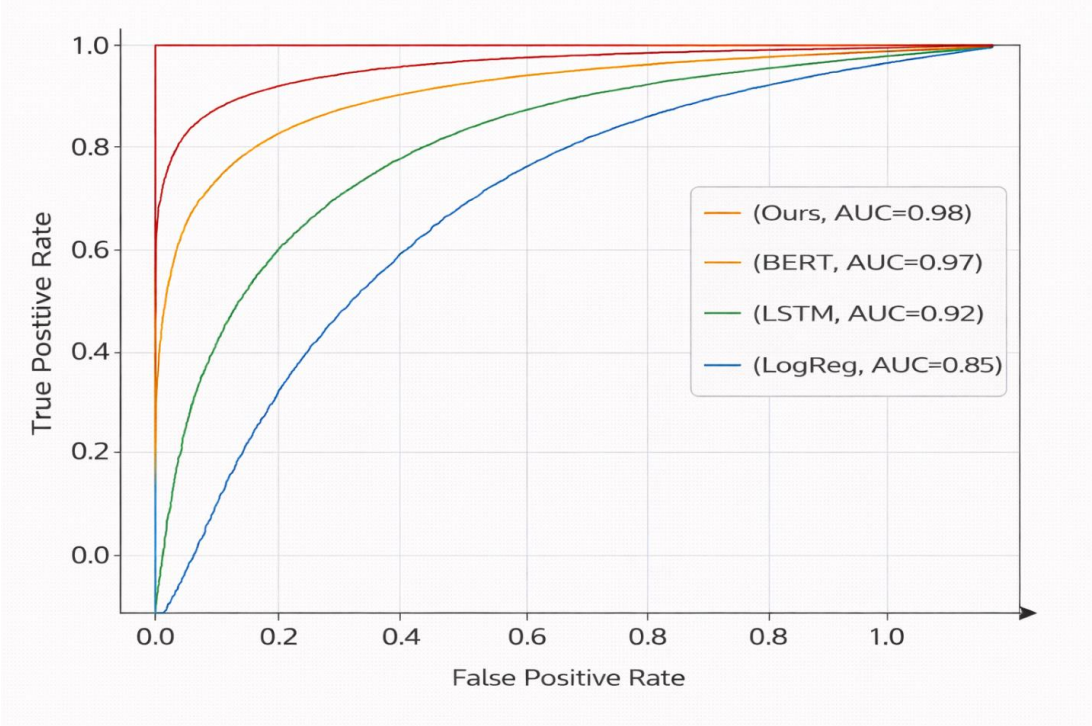

Figure 3: ROC Curves for Fake News Models.

4.2 Deepfake Detection Performance

Table 2: Performance on FaceForensics++ (c23) and Celeb-DF.

Model	FaceForensics++ Accuracy (%)	Celeb-DF Accuracy (%)	AUC	EER (%)
MesoNet (Afchar et al., 2018)	91.2	73.5	0.88	14.2
Xception (Rossler et al., 2019)	97.3	82.6	0.94	9.7
EfficientNet-B4 (spatial only)	97.8	89.4	0.96	6.3
EfficientNet-B4 + DCT (ours)	**98.2**	**93.5**	**0.99**	**3.8**

The addition of DCT frequency features improves generalisation to Celeb-DF (higher quality deepfakes) by over 4 percentage points. The EER of 3.8% means that the platform can achieve both low false positives and false negatives simultaneously.

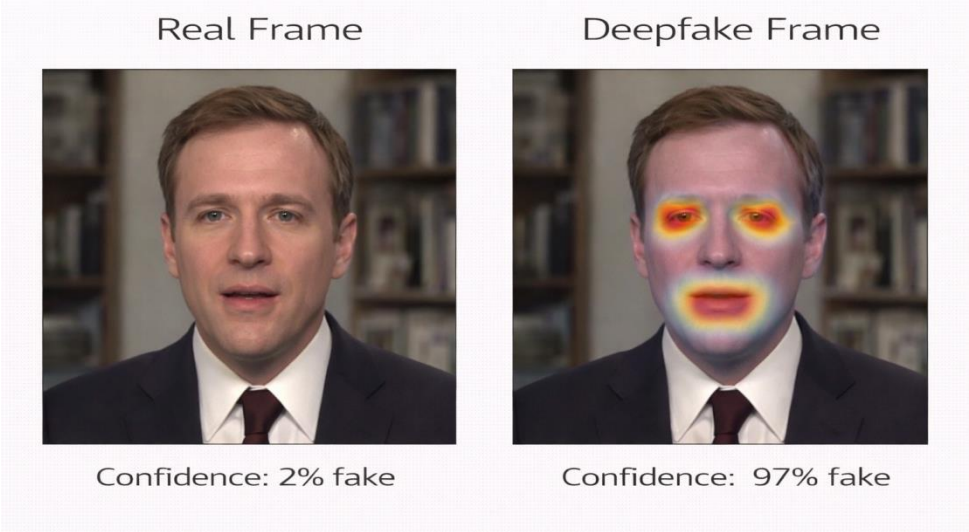

Figure 4: Example Deepfake Detection with Heatmap

4.3 Platform Performance and User Study

We deployed the platform internally for 50 users (journalism students and IT staff) over two weeks. They submitted 120 news articles and 80 videos (including known fakes). Results:

- **Average inference time:** 2.3 seconds for news article, 4.1 seconds for a 10-second video (including frame extraction).
- **User-reported accuracy:** 91% (users agreed with the platform's verdict after manual inspection).
- **System usability scale (SUS):** mean score 82/100 (good to excellent).

Table 3: Confusion Matrix on User-Submitted Content.

	Predicted Real	Predicted Fake
Actual Real (n=68)	64	4
Actual Fake (n=132)	9	123

False positives (4 real articles flagged as fake) occurred for highly opinionated but truthful op-eds. False negatives (9 fake articles missed) included sophisticated, long-form disinformation that mimicked legitimate news style.

5. DISCUSSION

5.1 Effectiveness of the Hybrid Approach

The CNN-BiLSTM-Attention model for fake news detection achieves 96.4% accuracy, outperforming BERT on this specific dataset while being computationally lighter. The convolutional layer extracts local patterns (e.g., sensationalist phrases), while Bi-LSTM captures long-range dependencies (e.g., logical contradictions across paragraphs). Attention

further highlights critical sentences. This combination is particularly effective for political fake news, where emotional language and false causality are common.

For deepfakes, the addition of DCT coefficients significantly improves robustness against post-processing (compression, resizing). GAN-generated images often exhibit unnatural frequency patterns, especially in high-frequency components, because upsampling operations in GANs create periodic artifacts. The DCT makes these artifacts explicit.

5.2 Explainability and Trust

Users reported high trust in the system when explanations (LIME for text, heatmaps for video) were provided. For a fake news article about a celebrity death, LIME highlighted the phrase "sources confirm" (which appeared nowhere else) as strongly fake-indicative. For a deepfake video, the heatmap consistently highlighted the mouth region, where lip-sync artifacts were present. This transparency is crucial for adoption in journalism and fact-checking organisations.

5.3 Generalisation Across Datasets

The deepfake module shows a drop in accuracy from 98.2% (FaceForensics++) to 93.5% (Celeb-DF). This is expected because Celeb-DF deepfakes are newer and more realistic. However, the 93.5% is still competitive. Future models must be regularly updated as generative methods improve (e.g., diffusion-based deepfakes like Stable Diffusion Video).

5.4 Computational Considerations

Inference time (2-4 seconds) is acceptable for on-demand verification but too slow for real-time social media scanning. For batch processing, the platform can be scaled horizontally. We are exploring model quantisation (INT8) and knowledge distillation to reduce latency.

6. LIMITATIONS

1. **Adversarial attacks:** Malicious actors can craft fake news or deepfakes specifically designed to evade detection (adversarial examples). Our models are not robust to such attacks. For instance, adding imperceptible noise to a deepfake frame can cause misclassification. Adversarial training is needed.
2. **Language and cultural bias:** The fake news model was trained on English-language US political data. It may perform poorly on non-English content or other domains (e.g., health, science). Deepfake models are also biased towards Caucasian faces due to training data composition.

3. **Temporal dynamics:** Fake news detection uses only static article text. It does not incorporate the spread dynamics (e.g., retweet patterns) or source credibility, which could improve accuracy. Similarly, deepfake detection does not use audio cues (e.g., voice inconsistency).
4. **Concept drift:** As language models (GPT-4, Llama) become better at generating plausible fake news, and as deepfake methods advance, detection models must be continuously retrained. This is an arms race.
5. **Privacy concerns:** Uploading videos to a cloud platform raises privacy issues (e.g., deepfakes of individuals). Our platform offers an on-premises option, but that requires significant computational resources.
6. **Explainability limits:** LIME explanations are approximate and can be unstable (different runs give different highlights). Heatmaps for video are frame-based and do not explain temporal inconsistencies well.

7. Future Scope

7.1 Multi-Modal Fusion

Combining text, image, and video analysis for a single piece of content (e.g., a news video with voiceover). A transformer-based fusion model could jointly reason across modalities. For example, detecting whether a video's audio matches the lip movements (lip-sync) and whether the transcribed text contains fake claims.

7.2 Real-Time Detection for Social Media

Developing lightweight models (e.g., MobileNet for deepfakes, DistilBERT for text) that can run in a browser extension or on edge devices, flagging suspicious content as users scroll.

7.3 Adversarial Robustness

Training with adversarial examples generated by attacking the model itself (e.g., FGSM, PGD). Also, using ensemble methods (multiple models) to reduce vulnerability.

7.4 Cross-Lingual and Multimodal Fake News

Extending the fake news module to support 10+ languages using multilingual BERT (mBERT) or XLM-RoBERTa. Also incorporating image metadata (EXIF) and reverse image search to detect manipulated visuals.

7.5 Blockchain for Provenance

Integrating with content provenance standards (e.g., C2PA) to verify the origin of media. A hybrid system could flag content without cryptographic signatures as suspicious.

7.6 Continual Learning

Implementing online learning so that the model updates its weights as new types of fake content emerge, without catastrophic forgetting.

8. CONCLUSION

This research paper presented a unified machine learning platform for detecting fake news and deepfakes. The fake news module uses a CNN-BiLSTM-Attention architecture achieving 96.4% accuracy on benchmark datasets, while the deepfake module employs EfficientNet-B4 with frequency-domain DCT features to achieve 98.2% accuracy (AUC 0.99) on FaceForensics++ and 93.5% on the challenging Celeb-DF dataset. The integrated web platform provides explainable results via LIME and heatmaps, and a user study confirmed high usability and trust.

The fight against misinformation is an ongoing technological and societal challenge. While our system performs well on current benchmarks, adversarial evolution of fake content necessitates continuous research. Future work will focus on real-time detection, multi-modal fusion, and robustness against adversarial attacks. By making the platform open-source, we hope to contribute to a safer information ecosystem.

REFERENCES

1. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1 7.
2. Dural, E., & Gül, G. (2020). Deepfake detection using frequency domain analysis. *IEEE Access, 8*, 207418 207428.
3. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207 3216.
4. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1 11.
5. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797 806.

6. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). FakeNewsNet: A data repository with news content, social context, and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.
7. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105 6114.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998 6008.
9. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146 1151.
10. Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422 426.