https://doi.org/10.31436/ijpcc.v12i1.651

Cross-Media Fake Content Detection via Independent Deep Learning Classifiers

Iqbal Najihah binti Samsul Kamal, Anna Safiya binti Samsudin, Raini binti Hassan*
Department of Computer Science, International Islamic University Malaysia, Gombak, Malaysia.

*Corresponding author: hrai@iium.edu.my
(Received: 15th December 2025; Accepted: 13th January, 2026; Published on-line: 30th January, 2026)

Abstract — The rapid advancement of generative models has enabled the creation of highly realistic fake multimedia content, including altered images, deepfake videos, and synthetic audio. These forgeries undermine information integrity and pose significant societal risks, especially by encouraging misinformation, digital fraud and impersonation. As these threats directly affect public trust and institutional transparency, they challenge the goals outlined in SDG 16: Peace, Justice, and Strong Institutions, which focuses on reducing corruption, preserving information integrity, and ensuring accountable, trustworthy systems. To address these issues, this paper proposes a deep learning–based system that classifies multimedia content across three modalities, which are image, video, and audio. Unlike conventional multimodal fusion approaches that necessitate paired data inputs, this paper introduces a novel routing-based unification architecture. The suggested framework makes use of a content-adaptive routing mechanism that treats each modality independently. Using a dual-backbone Swin Transformer and EfficientNet for images, Video Swin Transformer for video, and Wav2Vec 2.0 for audio, the system automatically determines the type of input file and sends it to the relevant specialized deep learning classifier. This design allows for a versatile, single-entry-point forensic tool that maintains high accuracy by leveraging domain-specific experts without the computational overhead of processing multiple streams concurrently. Experimental results demonstrate strong performance across individual modalities, with the audio model achieving 96.95% accuracy and the image model showing robust precision despite challenges posed by high quality generative forgeries.

Keywords— Deep Learning, Multimedia Forensics, Swin Transformer, Wav2Vec 2.0, Machine Learning, Data Science.

I. Introduction

The rapid development of generative models and artificial intelligence (AI) has drastically changed how multimedia content is created and altered. While these technologies have driven innovation in visual effects, digital media production, and human-computer interaction, they have also promoted the emergence of deepfakes, a highly realistic synthetic images, videos, and audio. Reliable detection methods are desperately needed because such content has the potential to disseminate false information, impersonate people, and weaken public trust in digital communication.

Deepfake techniques, including GANs, encoder-decoder models, and diffusion models, produce synthetic media that closely resembles real content, making manual detection more challenging [1]. These deepfakes have been used in identity fraud, political manipulation, disinformation campaigns, and various forms of social engineering, raising concerns for people, organizations, and public institutions. Consequently, the research community has prioritized developing automated systems that can distinguish between real and altered multimedia content across various modalities.

Although high-performance detection algorithms are available, there is still a big usability and system integration gap. Currently, many cutting-edge detection models are extremely specialized and made to handle a single modality, such as independently analysing only images, only video or only audio. Forensic analysts and regular users, who frequently need different software tools or platforms to verify various file types, are left with a fragmented landscape as a result. For instance, to verify a suspicious news report, it might be necessary to use one tool to look at the headline image and another environment to examine an audio clip that goes with it. This lack of unification slows down the reaction to disinformation campaigns and causes friction in the verification process.

This paper supports a unified "Cross-Media Fake Content Detection" framework using separate deep learning classifiers to address this fragmentation. This method emphasizes adaptability and architectural independence by developing core innovation of routing architecture that serves as a single interface for various media formats. The system cleverly directs the input to the most competent independent deep learning model by examining its structure.

To ensure the resilience of these independent classifiers, this paper makes use of a variety of modality-specific benchmark datasets. For images, the IMD2020 dataset offers a balanced set of real and manipulated samples involving inpainting and real-world forgeries [2], while the CASIA 2.0 dataset offers traditional image modifications like splicing, and copy-move editing. In the video domain, the DeeperForensic1.0 (v2) dataset, which contains complex face-swapping manipulations, serves as a high-quality benchmark for deepfake detection [3]. For audio, the ASVspoof dataset provides standardized real and spoofed speech samples, including synthetized voice-converted and synthetized audio [4]. Furthermore, the models of images, video, and audio are trained on distinct repositories to ensures that the specific artifacts unique to each medium are learned accurately. This data-driven strategy ensures that the system is accurate in its detection capabilities across various content types and unified in its interface.

In the end, this methodology guarantees that a user can confirm a suspicious file in a single streamlined environment, regardless of whether it is an image, voice recording, or video clip. This project intends to provide a detection tool that is both accurate and practically deployable for real-world scenarios where the format of the incoming threat is unpredictable by combining specialized, independent classifiers under a single "Cross-Media" umbrella. It is crucial to note that the developed framework is not a fully deployable commercial forensic tool, but rather a research-oriented prototype evaluated under controlled conditions to show the viability of cross-media routing.

II. RELATED WORK

This paper builds on several works that remain relevant in today's multimedia fake news landscape. Transformer-based models have become a strong foundation due to their pretraining on large, modern datasets which reduce the need for traditional handcrafted feature engineering.

For image forgery detection, this paper adopts a hybrid approach using Swin Transformer and EfficientNet, chosen for their ability to capture both global and fine-grained details. Prior work has tended tothis stufy lean toward one side of this spectrum, which creates a clear comparative context for the present aproach. B. Singh et al. (2022) [9] relied on EfficientNet-B0 within a multimodal setting, pairing it with a text encoder for credibility analysis. Their reliance on EfficientNet provided strong local feature extraction, but their fusion design predated transformer-based attention mechanisms. Compared with that framework, the current study benefits from Swin Transformer's hierarchical global reasoning, providing a broader contextual understanding that EfficientNet alone could not capture in Singh et al.'s setup.

Similarly, Almsrahad et al. (2024) [10] used EfficientNet-B0, though they focused on ELA-processed images from CASIA. Their results highlight EfficientNet's usefulness for low-level forensic patterns, yet their dependence on ELA artifacts limits robustness to modern social-media imagery. In contrast, this paper avoids hand-crafted preprocessing and instead integrates EfficientNet with Swin Transformer to balance low-level artifact detection with higher-level semantic consistency, addressing the brittleness seen in ELA-driven pipelines.

More recent work has leaned toward transformer-only designs. Gong et al. (2024) [11] applied Swin Transformer to video frames, introducing consistency-loss mechanisms to strengthen temporal generalization. Their focus on temporal cues, however, leaves open the question of how Swin could be paired with CNN-based forensic extractors. Mishra et al. (2023) [12] further showed that Swin outperforms many CNN baselines in robustness, but their evaluation—like other Swin-centric studies—prioritizes transformer capacity over hybrid feature diversity.

Across these studies, the pattern is clear where EfficientNet-based approaches excel at localized artifacts but struggle with global context, while Swin-based approaches capture global structure yet often overlook lo-level forensic detail. By combining both, this paper positions itself between the two extremes, aiming to inherit the strengths of each and mitigate their individual weaknesses.

For audio forgery detection, this paper uses a hybrid of wav2vec2, BiLSTM, and an attention mechanism to balance high-level speech representations with temporal modelling. Prior work by J. M. Martin-Donas et al. [13] established wav2vec2 as a strong front-end feature extractor for audio deepfake detection. While their model combined wav2vec2 with downstream classifiers, architectures integrating BiLSTM with attention were not explored, leaving a gap in modelling longer temporal dependencies as well as localized acoustic cues.

Samia et al. (2024) [14] explored a hybrid architecture of CNN, BiLSTM, and Multi-Head Attention, showing that combining temporal modeling with attention significantly boosts reliability. A key limitation, however, is their use of CNN-based spectral features, which restricts the model to handcrafted inputs. We address this by employing wav2vec2 to work directly with raw waveform representations. This approach leverages learned speech embeddings rather than static spectral cues. Ultimately, by coupling wav2vec2 with BiLSTM and attention, our model captures both global patterns and local anomalies more effectively.

For video forgery detection, this paper employs Video Swin Transformer as a standalone backbone, prioritizing its capacity to learn complex spatio-temporal patterns directly from raw video clips. This produces a more robust representation of motion-based manipulations compared to models that focus only on spatial cues. In contrast, Khalid et al. (2023) [15] used a Swin Y-Net Transformer, where the Y-Net design fused multi-scale features through parallel Swin branches. Their model effectively captured both local and global forgery signals, yet the limited dataset in their study introduced overfitting, especially for specific manipulation types. This restricts the generalizability that our Video Swin implementation aims to preserve through more balanced and diverse training data.

Deressa Zhou et al. (2023) [16] explored a different angle by combining ConvNeXt, Swin Transformer, and AE/VAE components to detect visual artifacts and latent inconsistencies. Their hybrid design improved generalization on unseen deepfake datasets thanks to the latent reconstruction loss. However, their approach remained frame-level and lacked a dedicated temporal modeling head, meaning it could not fully exploit motion cues. The method also depended heavily on precise face extraction; performance degraded noticeably when evaluated on full-frame inputs. In contrast, the current study avoids such dependency by using Video Swin's native spatio-temporal processing, reducing reliance on face cropping and allowing the model to handle a wider range of video structures.

Broadly, this paper unifies image, audio, and video detection under a single framework to address the fragmentation in forensic tooling. Although cross-modal analysis has been studied in the past, these studies frequently suffer from dependency on paired inputs. In 2022, Zhou et al. [17] applied CLIP to align image and text features, improving fake-news detection on datasets such as Weibo, PolitiFact, and GossipCop. Such fusion-dependent architectures work well for news articles that contain both, but they fall short when analysing isolated media files (such as a standalone audio recording or video clip) in the absence of related text.

Two years later, Ma et al. (2024) [18] proposed an event-aware multi-view fusion framework combining text, image, and additional signals. Their model reduced ambiguity in mismatched news content by emphasizing event structure, which is beneficial for real-world news contexts. Nevertheless, the system is computationally demanding and less useful for general-purpose forensics where the context is unknown due to its reliance on event-level consistency.

This supports the credibility of our study, since we focus on a 'content-agnostic' system. Unlike these rigid fusion architectures, our study suggests a Cross-Media Routing Framework. Our method does not require simultaneous data inputs by treating each modality with a specialized, independent deep learning classifier. This gives the system a degree of flexibility that strict multimodal fusion models don't, ensuring that it works whether the user submits a single image, a voice recording, or a video file.

III. METHODOLGY

The procedure for developing the cross-media fake content detection framework is described in this section. The intelligent routing mechanism that unifies them comes after dataset preparation, preprocessing pipelines, model architecture, training methods, and evaluation metrics for each independent classifier.

A. Fake Image Detection

- *Dataset preparation*

A final custom dataset of 28,000 images was created by randomly selecting 14,000 samples per class using a fixed seed to ensure class balance. A tuple (image_path, label) was used to store each entry, with label 0 denoting real and label 1 denoting fake. To ensure strong generalization and avoid information leakage, the dataset was divided into 70% training, 20% validation, and 10% testing after being shuffled using sklearn.utils.shuffle.

- *Preprocessing*

Two preprocessing pipelines were designed. The training pipeline included extensive augmentation to improve robustness against a variety of manipulation techniques. The transformations included RandomResizedCrop, RandomHorizontalFlip, RandomRotation (±10°), ColorJitter, RandomPerspective, GaussianBlur, RandomErasing, additive noise via a Lambda transform, and ImageNet normalization. These augmentations aid in exposing the model to generative artifacts and texture irregularities that are commonly found in manufactured media [5].

The evaluation pipeline only used to resize to 224x224 pixels, tensor conversion, and ImageNet normalization to ensure consistent and unbiased testing conditions.

- *Model Architecture*

A dual-backbone architecture was employed to take advantage of complementary visual representations. The first backbone, a Swin Transformer, offers hierarchical global-local modelling for the purpose of detecting subtle deepfake artifacts. The second backbone, EfficientNet-B3, uses a compound scaling design to capture fine-grained texture irregularities. Both backbones were kept completely frozen throughout training to minimize overfitting and training time.

Let F_{img} and F_{eff} indicates the embeddings generated by EfficientNet-B3 and the Swin Transformer. The definition of the fused representation is:

$$F = [F_{\mathrm{img}} ; F_{\mathrm{eff}}],$$

where [;] denotes vector concatenation. This fused feature vector is subsequently transformed into a binary real-fake prediction by a multi-layer classifier.

- *Training*

The model was optimized using a cross-entropy objective with regularization and a cosine-annealing learning-rate schedule. To increase training efficiency, automatic mixed precision was employed. Based on validation performance, early stopping with a patience of five epochs was used to avoid overfitting.

- *Evaluation*

Performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and a thorough per-class classification report. These metrics are in line with accepted methods in research on deepfake detection

B. *Fake Video Detection*

- *Dataset preparation*

For the video-based fake multimedia detection experiment, this paper utilized the DeeperForensic1.0 dataset, a large-scale benchmark for face manipulation detection. The dataset consists of high-quality real videos featuring 100 professional actors and their corresponding AI-generated videos, created using an end-to end face swapping framework []. A curated video dataset was created by sampling 2,00 real and 2,00 fake videos using a fixed seed to maintain class balance. A tuple (video_path, label) was used to store each dataset entry, with 0 denoting real and 1 denoting fake. To maintain the class distribution, the dataset was divided into 80% training and 20% validation.

- *Preprocessing*

Videos were preprocessed by uniformly sampling 8 frames per video, resizing frames to 224x224 pixels, and normalizing them using ImageNet statistics. During training, frame-level augmentation such as RandomResizedCrop and RandomHorizontalFlip were applied to improve robustness against varying visual content. For evaluation, only resizing and normalization were applied to maintain consistency.

- *Model architecture*

The foundation for video feature extraction was a Swin 3D Transformer (Tiny) that had been pretrained. To convert the extracted embeddings to binary predictions, a linear layer was used in place of the original classification head. Let F_vst represent the embeddings generated by the Swin3D backbone. The final forecast is calculated as follows:

$$y = softmax(FC(F_{\mathrm{vst}}))$$

Where FC is the fully connected classification layer.

- *Training*

The model was trained using cross-entropy loss and the Adam optimizer with a learning rate of 1×10^{-4}. Because of memory limitations, the batch size was set to 5. The model with the lowest validation loss was saved as the last checkpoint, and early stopping was implemented based on validation loss. To maintain consistent input shapes, video padding and frame extraction were carefully handled during the ten epochs of training.

- *Evaluation*

The model was evaluated on the validation set using confusion matrix to assess per-class performance. The approach guarantees that the model retains generalization to unseen samples while learning discriminative temporal and spatial patterns suggestive of manipulated videos.

C. *Fake Audio Detection*

- *Dataset preparation*

The audio modality was developed using the ASVspoof 2019 Logical Access (LA) corpus, which contains of bonafide human speech and spoofed utterances generated through various text-to-speech and voice conversion systems [4]. To ensure a controlled and balanced training set, all 2,580 bonafide samples were kept and matched with 2,580 randomly chosen spoofed samples using a fixed seed. For the development and evaluation subsets, stratified sampling was then used to reduce both subsets while maintaining the initial class imbalance, yielding in 410 bonafide and 3,590 spoof files for development and 413 bonafide and 3,587 spoof files for evaluation. A clean, organized dataset appropriate for representation learning and subsequent classification is created by pairing each audio file with its matching label (0=bonafide, 1=spoof).

- *Preprocessing*

The pretrained Wav2Vec 2.0 model, which offers self-supervised embeddings that capture phonetic, spectral, and prosodic cues pertinent to spoof detection, was used to transform all audio files into fixed-length feature representations. The extracted embeddings were stored as PyTorch tensors to ensure consistent input dimensionality and prevent repeated computation. Since audio duration varies among utterances, sequences were only padded at the batch level during loading, allowing the model to process variable-length speech while maintaining temporal patterns.

- *Model Architecture*

The classifier consists of a bidirectional LSTM and an attention mechanism that highlights the most informative temporal frames for differentiating between bonafide and spoofed speech. The bidirectional design allows the model to capture long-rage temporal dependencies, while the attention layer generated a weighted representation that concentrates on segments with spoof-related artifacts. The aggregated representation is mapped to a binary output (bonafide vs. spood) by a fully connected layer.

- *Training*

The model was trained using cross-entropy loss and the Adam optimizer with a fixed learning rate. Training proceeded for a limited number of epochs, and the final model was chosen based on the lowest validation loss to reduce overfitting. Since the dataset contains class imbalance in the development and evaluation sets, metrics were tracked across both classes to guarantee stable generalization.

- *Evaluation*

Accuracy, precision, recall, F1-score, and confusion matrices were used to evaluate the model's performance, enabling a comprehensive understanding of both bonafide and spoof classes. This evaluation framework provides insight into false-accept and false-reject tendencies, which are crucial in anti-spoofing applications. The modular design also ensures that the audio classifier can be easily incorporated into the entire multimodal late-fusion pipeline.

D. *Cross-Media Routing and System Integration*

To operationalize the separate classifiers into a single, coherent framework, a unified inference class was created using PyTorch. By acting as an intelligent router, this system shields the user from the intricacies of the underlying model.

- *System Initialization and Resource Management*

All three pre-trained model architectures are loaded into GPU memory (cuda) by the system upon instantiation. The specific weights for each classifier are loaded from independent .pth checkpoints. By maintaining these as separate files, the system allows for the individual updating of a particular modality without necessitating a full system retraining.

- *Intelligent Input Routing*

A routing mechanism based on file extension is used in the core logic. The system examines the extension when a file path is passed to the predict() function to identify the proper processing stream. If an unspoorted format is detected, the system raises an error, enduring processing stability.

- *Dynamic Preprocessing*

The inference pipeline places more emphasis on consistency than training pipelines, which heavily rely on augmentation. To ensure that input tensors match the dimensions required by the corresponding backbones, the system uses OpenCV (cv2) to sample fixed video frames, torchaudio to normalize audio sampling rates, and PIL to resize images.

- *Unified Output Standardization*

The system creates a probability distribution by passing the raw model logits through a Softmax layer, regardless of the modality employed. Three essential metrics are included in the final output, which are the modality employed, the prediction label, and a confidence score.

IV. RESULTS

A. *Fake Image Detection*

The experimental results for fake image detection is presented in Table. 1. Similarly, Figure 1 shows that the model performed well on the held-out test set consisting of 2,800 images. Overall, the model achieved an accuracy of 73.4%, precision 72.0%, recall of 76.3%, and an F1-score of 74.1%, proving a balanced performance in detecting real and fake images.

TABLE I
RESULT OF FAKE IMAGE DETECTION ON THE EVALUATION SET.

Metric	Score (%)
Accuracy	73.4
Precision	72.0
Recall	76.3
F1-score	74.1

According to per-class results (Table 2) and confusion matrix (Figure 1), the model classified 990 as real (70.5%) out of 1,405 real images, while 415 were misclassified as fake. Conversely, 1,065 images were accurately detected as fake (76.3%) among 1,395 fake images, whereas 330 images were incorrectly labelled as real. This illustrates the model's comparatively better ability to identify phony images, probably because of unique generation artifacts that are still present in contemporary synthetic image pipelines.

TABLE II
CLASSIFICATION REPORT (PER-CLASS) FOR FAKE IMAGE DETECTION ON THE EVALUATION SET

	Precision (%)	Recall (%)	F1-score (%)
Real (0)	75.0	70.5	72.7
Fake (1)	72.0	76.3	74.1
Accuracy	73.4	73.4	73.4

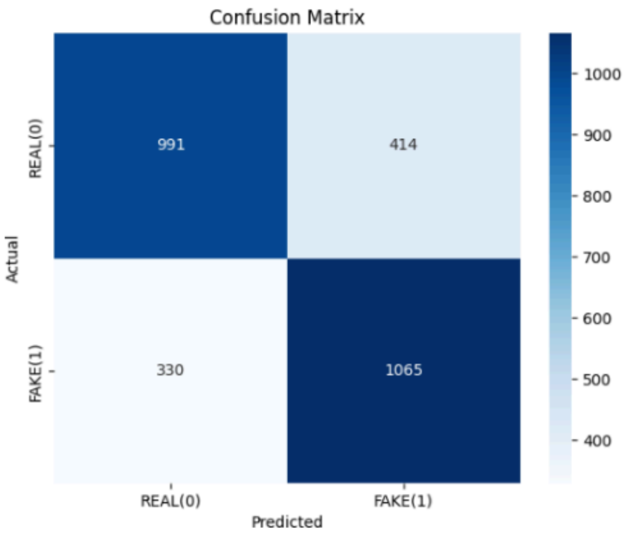


Fig. 1 Confusion Matrix for fake image detection

The model's discriminative ability is further supported by the ROC curve in Figure 2, which shows strong separability between the real and fake classes with a ROC-AUC of 0.824. The high AUC implies that the features successfully improve the model's capacity to discern minute cues present in altered images.

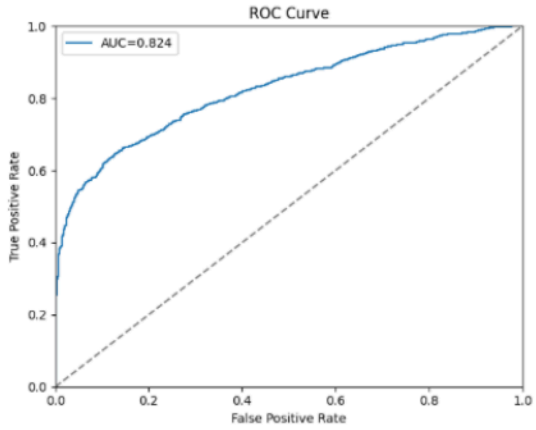


Fig. 2 ROC Curve for fake image detection

B. *Fake Video Detection*

Based on Table 3, an independent test set of 800 video samples, which have 400 real and 400 fake, was used to assess the video deepfake detection model. The model achieved 100% accuracy, precision, recall, and F1-score.

TABLE III
RESULT OF FAKE VIDEO DETECTION ON THE EVALUATION SET

Metric	**Score (%)**
Accuracy	100
Precision	100
Recall	100
F1-score	100

The per-class classification report shows perfect performance, with 100% precision, recall, and F1-score for both real and fake videos, yielding an overall evaluation accuracy of 100% (see Table 4).

TABLE IV
CLASSIFICATION REPORT (PER-CLASS) FOR FAKE VIDEO DETECTION ON THE EVALUATION SET.

	Precision (%)	**Recall (%)**	**F1-score (%)**
Real (0)	100	100	100
Fake (1)	100	100	100
Accuracy			100

According to Figure 3, a flawless classification pattern can be seen. All 400 of real videos were correctly classified as real (100%), with zero instances mistakenly identified as fake. Similarly, the model achieved a perfect score for fake videos (100%), correctly identifying each of the 400 instances with no false negatives. The absence of both false negative and false positive shows that the model maintains maximum specificity and sensitivity. Furthermore, the balanced precision and recall across classes proves that the model is unbiased toward either the real or fake class.

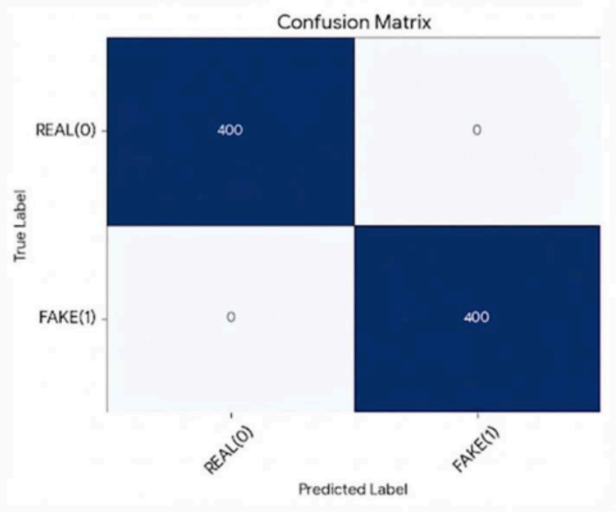


Fig. 3 Confusion Matrix for fake video detection

The ROC curve in Figure 4 shows perfect class separability with a ROC-AUC of 1.000, further supports the model's discriminative ability. This implies that complex spatiotemporal anomalies and synthesis artifacts present in phony videos are successfully captures by the Video Swin Transformer, enabling a clear differentiation between real and fake content.

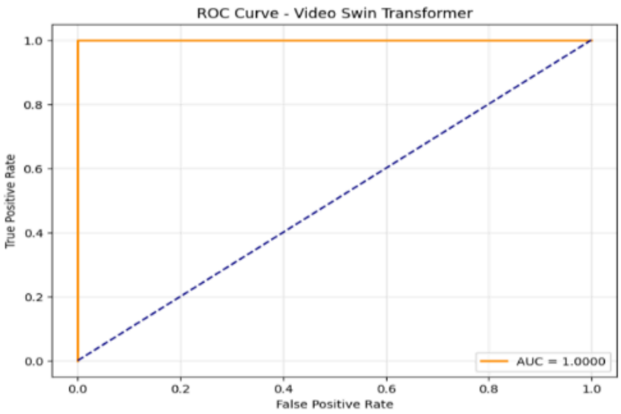


Fig. 4 ROC Curve for fake video detection

C. *Fake Audio Detection*

Fake audio detection achieved strong performance on the evaluation set. The model recorded an accuracy of 92.2%, demonstrating reliable overall classification. High precision of 99.4% indicates minimal false positives, while a recall of 91.8% reflects effective identification of fake audio samples (see Table 5). The F1-score of 95.5% confirms a balanced and robust detection capability, highlighting the model's effectiveness in distinguishing authentic and manipulated audio content under realistic evaluation conditions.

TABLE V
FINAL EVALUATION METRICS FOR FAKE AUDIO DETECTION ON THE EVALUATION SET.

Metric	**Score (%)**
Accuracy	92.2
Precision	99.4
Recall	91.8
F1-score	95.5

The model successfully identified 392 real samples (94.9%) out of 413, while 21 were incorrectly classified as fake, according to the per-class performance displayed in Tale 6 and the confusion matrix in Figure 5. On the other hand, out of 3,587 fake samples, the model correctly identified 3,294 (91.8%) of them, with 293 being mistakenly classified as real. This finding shows that the model is strong in detecting fake audio, as reflected in the very high precision, indicating that when the model predicts an audio clip as fake, it is almost always correct. Because of the inherent variability in human speech, real audio is still more difficult to model, as indicated by the comparatively lower recall for the real class.

TABLE IV
CLASSIFICATION REPORT (PER-CLASS) FOR FAKE AUDIO DETECTION ON THE EVALUATION SET.

	Precision (%)	**Recall (%)**	**F1-score (%)**
Real (0)	57.0	95.0	71.0
Fake (1)	99.0	92.0	95.0
Accuracy			92.0

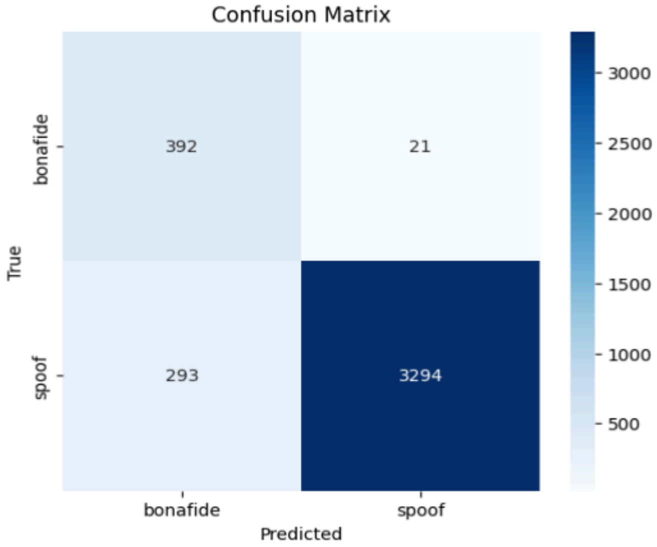


Fig. 5 Confusion matrix for fake audio detection.

The ROC curve in Figure 6, which shows a high level of class separability with a ROC-AUC of 0.972, further supports the model's discriminative ability. This suggests that the extracted Wav2Vec 2.0 embeddings effectively capture subtle acoustic inconsistencies and synthesis artifacts found in fake audio when combined with the BiLSTM and attention mechanism.

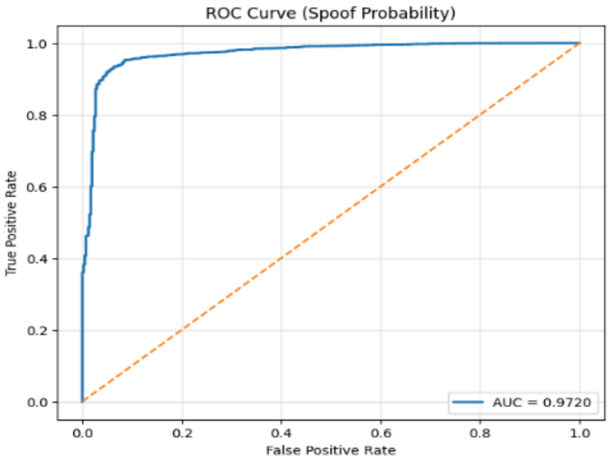
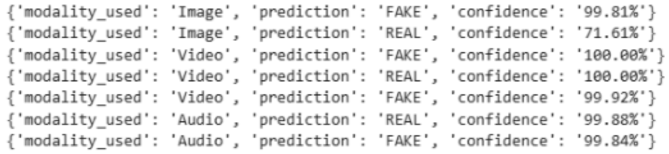


Fig. 6 ROC Curve for fake audio detection

D. System Integration and Cross-Media Verification

The fully integrated class was tested on a set of seven random samples that included a variety of media formats in order to verify the efficacy of the suggested routing architecture. To test robustness, the test set contained real-world media and benchmark samples from IMD2020, CASIA, DeeperForensic1.0, and ASVspoof.



```
{'modality_used': 'Image', 'prediction': 'FAKE', 'confidence': '99.81%'}
{'modality_used': 'Image', 'prediction': 'REAL', 'confidence': '71.61%'}
{'modality_used': 'Video', 'prediction': 'FAKE', 'confidence': '100.00%'}
{'modality_used': 'Video', 'prediction': 'REAL', 'confidence': '100.00%'}
{'modality_used': 'Video', 'prediction': 'FAKE', 'confidence': '99.92%'}
{'modality_used': 'Audio', 'prediction': 'REAL', 'confidence': '99.88%'}
{'modality_used': 'Audio', 'prediction': 'FAKE', 'confidence': '99.84%'}
```

Fig. 7 Sample predictions of multimodal detection

Based on Figure 7, the system successfully routed input files to the appropriate modality. This demonstrates that the predict() function's logic is dependable for mixed-media workflows. Across all modalities, the integrated system showed high levels of confidence. While synthetic content was consistently detected with confidence scores exceeding 99%, the lowest confidence recorded for a real image was 71.61%.

V. DISCUSSION

Cross-media fake news detection matters because misinformation spreads quickly and can destabilize communities. False content often carries emotional charge, creates confusion, and fuels misleading narratives that people may unknowingly amplify. Systems that can detect manipulated or misleading content across multiple modalities help reduce that risk and support healthier information ecosystems.

This paper employs a Content-Adaptive Routing Framework to tackle the problem of various multimedia forgeries. Our system operates as a unified forensic interface, in contrast to inflexible multimodal systems that require the integration of disparate data stream which often failing when a user provides only one file type. The input format (image, video, or audio) is dynamically identified by the system's routing logic, which then sends it to a specialized "expert" deep learning classifier. The system can successfully verify isolated media files without relying on paired data (e.g., requiring audio to accompany video) in part to this strategy's high availability and robustness.

Compared to monolithic fusion models, the suggested routing architecture has several engineering advantages. Because resources are only allocated to the appropriate model for a given input (for example, the heavy Video Swin model is never loaded into memory when analysing a simple JPEG), it is producing a computationally efficient system. Additionally, the design is very modular; future enhancements to the audio component, for instance, can be incorporated without requiring a full retraining of the image or video subsystems by simply updating the AudioModel class weights. The framework is a workable, scalable solution for real-world multimedia verification because of its flexibility.

A system like ours could be applied during elections, integrated into newsroom verification pipelines, or used in social-media monitoring to flag suspicious content before it gains traction.

Despite the strengths, there are limitations. Our project relies on publicly available datasets, which are relatively small and may not capture the full diversity of real-world social-media content. A larger, more varied dataset would improve generalization. Computational cost is another constraint: multimodal deep learning requires significant processing power, which can make experimentation slower and deployment more expensive.

VI. CONCLUSION

Cross-media using images, audio, and video provides a valuable technological approach for helping users avoid becoming victims of false information. This paper achieved strong performance across all three modalities, particularly in detecting *fake* content, which is typically more challenging. However, several limitations were encountered. The datasets were sourced from publicly available repositories, which may not be as current or diverse as datasets from private domains. This limits real-world representativeness. In addition, computational constraints due to budget limitations restricted the scale and complexity of the experiments. Future work can focus on reducing domain shift between controlled, lab-based datasets and real-world multimedia. Enhancing generalization in this way will help the model produce more robust results and better align with real-time, real-world data.

ACKNOWLEDGMENT

In the name of Allah, the Most Merciful and Gracious. We are grateful to Allah (SWT) for giving us the knowledge and perseverance to finish this paper. For their steadfast support, we are grateful to our parents, instructors, peers, and supervisor. May this endeavour gain Allah's blessings and benefit society. Finally, the authors would like to thank the Department of Computer Science at IIUM for supporting this paper.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

I.N. Samsul Kamal and A.S. Samsudin contributed to the design and implementation R. Hassan provided supervision, validated the methodology, and reviewed the final manuscript.

DATA AVAILABILITY STATEMENT

All datasets utilized in this paper are sourced from publicly available repositories. The IMD2020, CASIA 2.0, DeeperForensic1.0 and ASVspoof 2019 datasets can be accessed via their respective citations.

ETHICS STATEMENT

This paper utilized exclusively publicly available benchmark datasets. No private data was collected, and no human subjects or animals were involved in the experimentation process.

REFERENCES

[1]. L. Verdoliva, "Media Forensics and DeepFakes: an overview," *arXiv preprint, arXiv:2001.06564*, Jan. 2020. [Online]. Available: https://arxiv.org/pdf/2001.06564

[2]. A. Novozámský, B. Mahdian, and S. Saic, "IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Snowmass Village, CO, USA, Mar. 2020, pp. 71-80, doi: 10.1109/WACVW50321.2020.9096940

[3]. *EndlessSora/DeeperForensics-1.0: [CVPR 2020] A Large-Scale Dataset for Real-World Face Forgery Detection.* (2025). GitHub. https://github.com/EndlessSora/DeeperForensics-1.0

[4]. H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, Sept. 2021. [Online]. Available: https://arxiv.org/abs/2109.00535.

[5]. B. Dolhansky *et al.*, "The DeepFake Detection Challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.07397

[6]. Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv preprint arXiv:2103.14030*, 2021. [Online]. Available: http://arxiv.org/abs/2103.14030

[7]. *EndlessSora/DeeperForensics-1.0: [CVPR 2020] A Large-Scale Dataset for Real-World Face Forgery Detection.* (2025). GitHub. https://github.com/EndlessSora/DeeperForensics-1.0

[8]. Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[9]. B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21503–21517, 2021. https://doi.org/10.1007/s00521-021-06086-4

[10]. Y. Almsrahad and N. M. Charkari, "Image Fake News Detection using Efficient NetB0 Model," *Journal of Information Systems and Telecommunication (JIST)*, vol. 12, no. 45, pp. 41–48, 2024. https://doi.org/10.61186/jist.40976.12.45.41

[11]. L. Y. Gong, X. J. Li, and P. H. J. Chong, "Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector," *Electronics*, vol. 13, no. 15, p. 3045, 2024. https://doi.org/10.3390/electronics13153045

[12]. S. R. Mishra, H. Mohapatra, S. A. Edalatpanah, and M. K. Gourisaria, "Advanced deepfake detection leveraging swin transformer technology," *Engineering Review*, vol. 44, no. 4, pp. 45–56, 2024. https://doi.org/10.30765/er.2583

[13]. J. M. Martín-Doñas and A. Álvarez, "The Vicomtech Audio Deepfake Detection System based on Wav2Vec2 for the 2022 ADD Challenge," *arXiv preprint arXiv:2203.01573*, 2022. [Online]. Available: https://arxiv.org/abs/2203.01573

[14]. S. Dilbar, M. A. Qureshi, S. K. Noon, and A. Mannan, "AudioFakeNet: A Model for Reliable Speaker Verification in Deepfake Audio," *Algorithms*, vol. 18, no. 11, p. 716, 2025. https://doi.org/10.3390/a18110716

[15]. F. Khalid, M. H. Akbar, and S. Gul, "SWYNT: Swin Y-Net Transformers for Deepfake Detection," in *2023 International Conference on Robotics and Artificial Intelligence (ICRAI)*, 2023, pp. 1–6. https://doi.org/10.1109/icrai57502.2023.10089585

[16]. Wodajo, D., Atnafu, S., & Akhtar, Z. (n.d.). "Deepfake Video Detection Using Generative Convolutional Vision Transformer," *arXiv preprint arXiv:2307.07036*, 2023. [Online]. Available: https://arxiv.org/pdf/2307.07036

[17]. Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang, "Multimodal Fake News Detection via CLIP-Guided Learning," *arXiv preprint arXiv:2205.14304*, 2022. [Online]. Available: https://arxiv.org/abs/2205.14304

[18]. Ma, Z., Luo, M., Guo, H., Zeng, Z., Hao, Y., & Zhao, X. (2024). "Event-Radar: Event-driven Multi-View Learning for Multimodal Fake News Detection," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5809–5821. https://doi.org/10.18653/v1/2024.acl-long.316

[19]. Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," *arXiv preprint arXiv:2106.13230*, 2021. [Online]. Available: https://arxiv.org/abs/2106.13230

[20]. Y. Sun, X. Li, J. Wang, L. He, and X. Liu, "Audio Anti-Spoofing Based on Audio Feature Fusion," *Algorithms*, vol. 16, no. 7, p. 317, 2023. https://doi.org/10.3390/a16070317

[21]. M. Li and X.-P. Zhang, "Interpretable Temporal Class Activation Representation for Audio Spoofing Detection," in *Interspeech 2024*, 2024. [Online]. Available: https://arxiv.org/abs/2406.08825

[22]. X. Liu, W. Ge, X. Wang, and J. Yamagishi, "LENS-DF: Deepfake Detection and Temporal Localization for Long-Form Noisy Speech," in *IJCB 2025*, 2025. [Online]. Available: https://arxiv.org/abs/2507.16220

[23]. Sivaraman, D. K., Saif, M., MR, M. F., & Moosa, M. (2025). Enhanced Fake Image Localization in Social Media using Swin Transformer and EfficientNet Feature Fusion. *International Journal for Research in Applied Science and Engineering Technology*, 13(4), 4052–4058. https://doi.org/10.22214/ijraset.2025.69194

[24]. S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: Analysing Model Generalisation Across Architectures, Datasets and Pre-Training Paradigms," *IEEE Access*, vol. 12, pp. 1880–1908, 2024. https://doi.org/10.1109/access.2023.3348450

[25]. A. Novozámský, B. Mahdian, and S. Saic, "IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2020, pp. 71–80. https://doi.org/10.1109/wacvw50321.2020.9096940

[26]. D. Goel, "CASIA 2.0 Image Tampering Detection Dataset," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/datasets/divg07/casia-20-image-tampering-detection-dataset

[27]. EndlessSora, "DeeperForensics-1.0," *GitHub repository*, 2025. [Online]. Available: https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/dataset

[28]. D. Wan, M. Cai, S. Peng, W. Qin, and L. Li, "Deepfake Detection Algorithm Based on Dual-Branch Data Augmentation and Modified Attention Mechanism," *Applied Sciences*, vol. 13, no. 14, p. 8313, 2023. *https://doi.org/10.3390/app13148313*