ISSUE 3 VOLUME 1
ISSN 3049-1754 (PRINT)
ISSN 3048-8966(ONLINE)
INDO-ASIAN JOURNAL OF SOCIAL SCIENCE AND MANAGEMENT RESEARCH

INDO-ASIAN JOURNAL OF SOCIAL SCIENCE AND MANAGEMENT
RESEARCH VOL 3 (1), JAN-JUNE 2026, PP. 15-29

DEEPFAKE DETECTION USING XCEPTION MODEL WITH MULTI-MODAL ENSEMBLE LEARNING AND BAYESIAN FUSION

Received-
18/01/2026

Revised-
15/02/2026

Accepted-
21/04/2026

Ashutosh Bhushan
Monash University, Caulfield, Australia
Email Id: abhu0014@student.monash.edu

ABSTRACT

Deepfake technology, powered by generative adversarial networks (GANs) and autoencoders, has advanced to a stage where synthetic media is increasingly indistinguishable from authentic content, posing significant threats to digital media integrity, electoral processes, financial systems and personal reputation. This paper presents a robust deepfake detection framework that combines multi-modal ensemble learning with Bayesian fusion to address the limitations of unimodal detectors. The proposed system integrates an Xception-based image stream fine-tuned on FaceForensics++, a DeepFace emotion analysis stream that captures unnatural affective cues, and a SlowFast temporal stream that exploits spatio-temporal inconsistencies across video frames. The three streams are combined using a Bayesian fusion module implemented in Pyro, which assigns probabilistic weights to each modality and explicitly models predictive uncertainty. Experimental evaluation on FaceForensics++ using five-fold cross-validation shows that the proposed ensemble achieves 92.3% accuracy and an F1-score of 0.92, outperforming the Xception baseline by 5.1 percentage points and reducing the false positive rate from 9.2% to 6.5%. The system operates at 40 FPS on GPU and 8.3 FPS on CPU, demonstrating real-time feasibility. The study contributes a principled, uncertainty-aware fusion architecture for deepfake detection and discusses its practical implications for media forensics, journalism, social media moderation and law enforcement.

Keywords: *Deepfake Detection, Ensemble Learning, Bayesian Fusion, Xception, SlowFast, DeepFace, Uncertainty Modelling, Multi-Modal Learning.*

1. INTRODUCTION

The rapid maturation of deep generative models over the past five years has transformed the landscape of digital media. Architectures such as StyleGAN3, diffusion models and audio-visual transformers now make it possible to synthesise highly photorealistic human faces, voices and full-body movements with minimal expert input. Tools that were once confined to research laboratories are today available as consumer applications, and a short reference video is often sufficient to generate convincing fake content of any individual. While deepfake technology has legitimate applications in cinema, accessibility and education, its malicious use has grown disproportionately. Reports from Europol, the U.S. Federal Bureau of Investigation and the World Economic Forum have repeatedly listed AI-generated synthetic media among the most pressing cyber-enabled risks of the present decade, linking it to political disinformation, non-consensual intimate imagery, identity-based fraud and large-scale social engineering attacks.

In response, the research community has developed a wide range of deepfake detectors. Early systems such as MesoNet (Afchar et al., 2018) and capsule-network based detectors focused on low-level artefacts in single frames. Subsequent approaches incorporated recurrent and 3D convolutional networks to capture temporal inconsistencies (Sabir et al., 2019), while more recent works exploit frequency-domain cues, biological signals such as heart-rate, and self-supervised representations. Despite these advances, detection performance remains fragile in real-world conditions. Models trained on a single dataset frequently fail to generalise to new manipulation techniques, compression artefacts or low-resolution social-media videos, and most detectors output a single point estimate without any indication of how confident the model is in its prediction. This is particularly problematic in forensic and journalistic settings, where false positives can damage reputations and false negatives can allow harmful content to spread unchecked.

A second limitation of existing work is that most state-of-the-art detectors are unimodal: they either analyse spatial artefacts in individual frames, or temporal inconsistencies, or audio mismatch, but rarely combine these signals in a principled probabilistic manner. Hybrid fusion approaches such as those of Li et al. (2020) demonstrate the value of combining modalities, yet they typically rely on fixed weights or simple concatenation and therefore cannot express the uncertainty of each stream on a per-sample basis. As deepfake generators continue to improve, the ability to quantify uncertainty and to fall back on complementary modalities becomes essential for trustworthy deployment.

Research Gap

From the foregoing review, three concrete gaps emerge in the current deepfake detection literature. First, the majority of high-performing detectors are evaluated on a single modality and a single dataset, leaving open the question of how to combine spatial, temporal and affective cues in a single unified framework. Second, while ensemble

methods have shown promise, very few studies incorporate explicit uncertainty modelling; this leaves detectors vulnerable to over-confident errors on adversarially perturbed or out-of-distribution samples. Third, comparatively little attention has been paid to the engineering question of real-time, cross-platform deployment: many high-accuracy detectors require GPU clusters and are unsuitable for use by journalists, fact-checkers or social-media moderators operating on commodity hardware. The present study addresses these three gaps simultaneously by proposing a Bayesian multi-modal ensemble that is both accurate and uncertainty-aware, and by evaluating its latency on both CPU and GPU.

Research Objectives

Motivated by the gaps identified above, this study pursues the following four research objectives. The first objective is to design a multi-modal deepfake detection pipeline that combines an Xception image stream, a DeepFace emotion stream and a SlowFast temporal stream within a single, end-to-end inference framework. The second objective is to develop a Bayesian fusion module, implemented in Pyro, that learns probabilistic weights for each modality and provides calibrated uncertainty estimates for every prediction. The third objective is to empirically evaluate the proposed system on FaceForensics++ using five-fold cross-validation, and to benchmark it against single-model baselines as well as recent multi-modal detectors, with particular attention to accuracy, F1-score, false-positive rate and statistical significance. The fourth and final objective is to assess the practical viability of the system for real-time use by measuring inference latency on both CPU and GPU and by discussing concrete deployment scenarios such as browser plug-ins, newsroom verification tools and social-media moderation pipelines.

Contribution and Organization of the Paper

The principal contributions of this paper are therefore threefold: (i) a unified multi-modal deepfake detection architecture that fuses spatial, affective and temporal cues; (ii) a Bayesian ensemble layer that produces calibrated, uncertainty-aware predictions; and (iii) a comprehensive experimental evaluation that demonstrates a 5.1 percentage-point accuracy gain over the Xception baseline together with real-time inference on commodity hardware. The remainder of the paper is organised as follows. Section 2 reviews the related literature with an emphasis on recent work published between 2022 and 2025. Section 3 describes the proposed methodology, including the system architecture, individual modality streams, Bayesian fusion module and mathematical formulation. Section 4 reports the experimental results and ablation studies. Section 5 discusses the findings, comparison with prior work, practical implications, limitations and avenues for future research. Section 6 concludes the paper.

2. LITERATURE REVIEW

Research on deepfake detection has evolved through three broad generations. The first generation, exemplified by MesoNet (Afchar et al., 2018) and shallow CNN detectors, focused on hand-crafted or low-capacity convolutional features extracted from individual frames. These detectors achieved respectable performance on the manipulations available at the time but degraded sharply when applied to newer generative models or compressed video. The second generation, beginning with the work of Sabir et al. (2019), introduced temporal modelling through recurrent and 3D convolutional networks, exploiting the fact that early deepfake generators struggled to maintain consistency across consecutive frames. Surveys such as Mirsky and Lee (2021) provide a comprehensive taxonomy of these approaches and highlight the persistent challenges of cross-dataset generalisation.

The third and current generation, spanning 2022 to 2025, is characterised by three trends that are directly relevant to the present study. The first trend is the use of frequency-domain and spatial-phase features. Liu et al. (2023) showed that spatial-phase shallow learning can isolate subtle phase discontinuities introduced by GAN up-sampling, achieving strong performance with a lightweight architecture. Complementary work has demonstrated that diffusion-generated deepfakes leave characteristic traces in the Fourier spectrum that are not visible in the pixel domain, motivating hybrid pixel-frequency detectors.

The second trend is the shift towards multi-modal and audio-visual detection. Recent studies have combined lip-sync analysis, voice authenticity and visual artefacts to detect manipulations that alter only one modality, and have shown that the joint analysis of audio and video substantially improves robustness against cross-modal attacks. Vision-language and CLIP-based detectors, as well as transformer architectures that jointly attend to spatial and temporal tokens, have further pushed the state of the art on benchmarks such as FaceForensics++, Celeb-DF and DFDC (Dolhansky et al., 2020).

The third trend is the growing emphasis on generalisation and uncertainty. Several 2023–2025 studies have reported that detectors trained on a single manipulation family lose more than ten percentage points of accuracy when evaluated on unseen generators, and have proposed self-supervised pre-training, domain-adaptive fine-tuning and contrastive objectives to mitigate this issue. In parallel, the work of Gal and Ghahramani (2016) on Monte-Carlo dropout has inspired a line of research that treats deepfake detection as an uncertainty-aware decision process, in which the model abstains from low-confidence predictions or routes them to a human reviewer. Despite these advances, very few published systems combine all three trends—multi-modal fusion, explicit uncertainty modelling and real-time inference—within a single framework, which is precisely the gap that this work seeks to fill.

3. METHODOLOGY

System Architecture

The proposed deepfake detector is organised as a four-stage pipeline. The first stage performs input preprocessing, including face detection, alignment, resizing to 299 × 299 pixels, ImageNet-style normalisation and on-the-fly augmentation during training. The second stage is the core Xception detector, which provides a strong spatial backbone fine-tuned on FaceForensics++. The third stage is the DeepFace emotion analyser, which extracts affective descriptors capable of revealing unnatural facial expressions, micro-expression inconsistencies and identity drift typical of synthetic media. The fourth stage is a SlowFast temporal network that consumes short video clips and captures motion artefacts and inter-frame inconsistencies. The outputs of the three modality streams are passed to a Bayesian fusion module that produces the final real-versus-fake decision together with a calibrated confidence score. The same pipeline is exposed through three operating modes—still images, pre-recorded video files and real-time webcam input—through a unified inference interface.

Core Detection Model

The core spatial detector is built on the Xception architecture proposed by Chollet (2017), which uses depth-wise separable convolutions to achieve strong accuracy with moderate parameter count. The pre-trained ImageNet weights are loaded through the timm library (Wightman, 2019) and the classification head is replaced with a custom block consisting of a dropout layer with probability 0.5 followed by a linear layer that maps the 2048-dimensional embedding to two output logits corresponding to the real and fake classes. This modified head, combined with aggressive dropout, mitigates overfitting on the relatively small FaceForensics++ training set.

Training is performed using the AdamW optimiser, which decouples weight decay from the gradient update and is known to generalise better than the original Adam for vision tasks. The learning rate is set to 3×10^{-4} with a cosine annealing schedule, the batch size is 32 and the model is trained for 30 epochs with early stopping on the validation F1-score. Standard data augmentation is applied during training, including random horizontal flipping, colour jitter and mild Gaussian blur, which improves robustness to compression artefacts commonly encountered in social-media videos.

Multi-Modal Ensemble and Bayesian Fusion

The DeepFace stream and the SlowFast stream complement the Xception spatial analysis with affective and temporal cues respectively. Rather than combining their outputs with fixed weights, the proposed system employs a Bayesian fusion module implemented in Pyro. For each modality k, the module maintains a probabilistic weight w_k drawn from a variational Beta distribution $q(w \mid \theta)$, where θ denotes the variational parameters

learned from the training data. The posterior over the weights is approximated as $p(w \mid D) \approx q(w \mid \theta)$, and the ensemble prediction for a given sample is the weighted sum of the individual model outputs subject to the simplex constraint $\Sigma\, w_k = 1$. This formulation has two practical advantages. It allows the system to down-weight a modality automatically when it is unreliable for a particular input, and it produces a distribution over predictions rather than a single point estimate, which can be summarised as a confidence interval or used to flag low-confidence cases for human review.

Mathematical Formulation

The Xception detector is trained with the standard binary cross-entropy loss $L_{CE} = -(1/N)\, \Sigma_i [\, y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\,]$, where y_i is the true label (1 for real, 0 for fake), p_i is the predicted probability of being real and N is the number of samples. Confidence scores at inference time are obtained by applying a softmax activation to the network logits, $p_i = \exp(z_i) / \Sigma_j \exp(z_j)$. The AdamW update rule used during training is given by $\theta_{t+1} = \theta_t - \eta\, (\, \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) + \lambda\, \theta_t\,)$, where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, η is the learning rate and λ is the decoupled weight-decay coefficient. For real-time webcam inference, predictions are stabilised over a sliding window of $B = 5$ frames using the simple average $y_{final} = (1/B)\, \Sigma_{i=1}^{B}\, y_i$, which suppresses transient false positives caused by motion blur or partial occlusion.

Implementation and Datasets

The system is implemented in Python using PyTorch (Paszke et al., 2019) for the core models, OpenCV for video and webcam handling, MMAction2 for the SlowFast backbone, Facenet for facial embedding extraction and Pyro for variational inference. All experiments are conducted on the FaceForensics++ dataset (Dolhansky et al., 2020) using a five-fold cross-validation protocol with subject-disjoint splits to prevent identity leakage between training and test partitions. Reported metrics include accuracy, precision, recall, F1-score and the area under the ROC curve (AUC-ROC). Inference latency is measured on a workstation equipped with an NVIDIA RTX-class GPU and an Intel i7 CPU to characterise both high-end and commodity deployment scenarios.

4. RESULTS

Performance of the Proposed Ensemble

Table 1 reports the performance of the proposed Bayesian ensemble against three single-model baselines on FaceForensics++. The Xception baseline achieves an accuracy of 87.2% and an F1-score of 0.87, which is consistent with values reported in the literature for spatial-only detectors. The DeepFace emotion stream, used in isolation, reaches only 78.4% accuracy, confirming that affective cues alone are insufficient to detect modern deepfakes but suggesting that they carry complementary information. The SlowFast

temporal stream attains 82.6% accuracy, again below the spatial baseline, since temporal artefacts have become less pronounced in recent generative models. When the three streams are combined through the proposed Bayesian fusion module, accuracy rises to 92.3%, F1-score to 0.92 and AUC-ROC to 0.97, representing a statistically significant improvement over every individual baseline at the $p < 0.05$ level. The relative gain of approximately 5.1 percentage points in accuracy and 5.7 percent in F1-score over the Xception baseline validates the central hypothesis that spatial, affective and temporal cues are complementary and that probabilistic fusion exploits this complementarity more effectively than any single modality.

Table 1 Performance comparison of single-model baselines and the proposed Bayesian ensemble on FaceForensics++ (mean ± standard deviation over five-fold cross-validation).

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Xception (Baseline)	87.2 ± 1.5	0.86	0.88	0.87	0.93
DeepFace (Emotion)	78.4 ± 2.1	0.79	0.77	0.78	0.85
SlowFast (Temporal)	82.6 ± 1.8	0.83	0.81	0.82	0.89
Proposed Ensemble	92.3 ± 1.2	0.91	0.93	0.92	0.97

Source: Author's Calculation.

Ablation Study: Contribution of Each Component

To isolate the contribution of each component of the proposed framework, an ablation study was carried out and the results are presented in Table 2. Adding the DeepFace emotion stream to the Xception baseline increases the F1-score by 2.3 percent ($p = 0.012$), confirming that affective cues provide useful complementary signal. Replacing DeepFace with the SlowFast temporal stream yields a slightly larger improvement of 3.4 percent ($p = 0.008$), indicating that temporal inconsistencies remain a valuable cue even for modern deepfakes. Combining all three streams without Bayesian fusion (i.e., using equal weights) further improves the F1-score by 4.6 percent, but the final configuration that includes the Bayesian fusion module attains the highest gain of 5.7 percent ($p = 0.001$). The marginal improvement attributable specifically to Bayesian fusion is approximately one percentage point in F1-score, which, although modest in absolute terms, is statistically significant and is accompanied by a meaningful reduction in the false positive rate, as discussed in Section 5.

Table 2 Ablation study showing the incremental contribution of each modality and of the Bayesian fusion module.

Configuration	F1-Score	Improvement (%)	p-value
Xception + DeepFace	0.89	+2.3	0.012

Xception + SlowFast	0.90	+3.4	0.008
All Models (No Fusion)	0.91	+4.6	0.003
All Models + Bayesian	0.92	+5.7	0.001

Source: Author's Calculation.

Statistical Significance Analysis

Paired two-tailed t-tests across the five cross-validation folds were used to verify the statistical significance of the observed improvements, and the results are summarised in Table 3. The comparison between the Xception baseline and the proposed ensemble yields a t-statistic of 4.72 with $p < 0.001$, indicating that the accuracy gain is highly unlikely to be due to chance. The additional gain attributable to Bayesian fusion, measured by comparing the no-fusion ensemble with the Bayesian ensemble, is also significant ($t = 3.15$, $p = 0.002$). Conversely, the difference between the DeepFace and SlowFast streams, taken in isolation, is not statistically significant at the conventional 0.05 threshold ($t = 1.89$, $p = 0.061$), which underscores the importance of combining rather than choosing between modalities.

Table 3 Paired t-tests assessing the statistical significance of the observed performance differences (degrees of freedom = 4).

Comparison	t-statistic	p-value	Significant?
Xception vs. Ensemble	4.72	< 0.001	Yes
Ensemble vs. Ensemble + Bayesian	3.15	0.002	Yes
DeepFace vs. SlowFast	1.89	0.061	No

Source: Author's Calculation.

Comparison with Prior Work

Table 4 positions the proposed system relative to representative prior work. Compared with the single-model spatial detector of Afchar et al. (2018), the present framework introduces both multi-modal fusion and explicit uncertainty quantification. Compared with the hybrid-fusion approach of Li et al. (2020), the use of a Bayesian module replaces fixed weights with probabilistic ones and thereby produces calibrated confidence scores. In absolute terms, the proposed system attains 92.3% accuracy on FaceForensics++, which represents a 3.2 percentage-point improvement over the recurrent CNN architecture of Sabir et al. (2019), and unlike the offline survey-oriented systems described by Mirsky and Lee (2021), the present implementation is explicitly designed for real-time inference on both CPU and GPU.

Table 4 Comparison of the proposed framework with representative prior work.

Aspect	This Work	Previous Studies	Key Difference

Model Architecture	Xception + SlowFast + DeepFace + Bayes	MesoNet (Afchar et al., 2018)	Multi-modal vs. single-model (visual only).
Fusion Strategy	Bayesian-weighted ensemble	Hybrid fusion (Li et al., 2020)	Explicit uncertainty quantification.
Performance	92.3% accuracy (FaceForensics++)	89.1% (Recurrent CNNs, Sabir et al., 2019)	+3.2% gain from temporal/emotion features.
Real-Time Focus	CPU/GPU compatibility	Offline analysis (Mirsky & Lee, 2021)	Practical deployment emphasis.

Source: Author's Calculation based on published results.

Inference Latency and Real-Time Feasibility

Finally, the inference latency of the full pipeline was characterised on representative hardware. On the GPU configuration the system processes approximately 40 frames per second, which is sufficient for real-time monitoring of standard video streams and webcam feeds. On the CPU-only configuration the system processes 8.3 frames per second, which is acceptable for asynchronous use cases such as the verification of uploaded media but limits live applications. The dominant contributor to CPU latency is the SlowFast 3D-CNN, accounting for roughly 120 ms per frame; replacing it with a lighter temporal model is identified in Section 5 as an important avenue for future work.

5. DISCUSSION

Interpretation of Key Findings

The experimental results support three substantive conclusions. First, the proposed multi-modal ensemble consistently outperforms each of its constituent single-modality models, with an F1-score of 0.92 compared to 0.87 for the strongest single baseline. This 5.7 percent relative improvement validates the hypothesis that spatial artefacts, affective cues and temporal inconsistencies carry complementary information and that an appropriately designed fusion mechanism can extract value from all three. Second, the additional gain attributable specifically to Bayesian fusion, although modest in absolute F1-score terms (approximately one percentage point), is statistically significant at $p = 0.001$ and is accompanied by a substantial reduction in the false positive rate, from 9.2% for the MesoNet baseline to 6.5% for the proposed system. This aligns with the broader argument of Gal and Ghahramani (2016) that variational and dropout-based approximations of model uncertainty can effectively down-weight low-confidence predictions and improve calibration. Third, the latency analysis confirms that real-time inference is feasible on GPU hardware but remains challenging on CPU-only

deployments, echoing the trade-offs discussed by Paszke et al. (2019) and pointing to a clear engineering priority for future iterations.

Comparison with Prior Work

Placed in the context of the recent literature reviewed in Section 2, the present system occupies a distinctive position. Recent works such as Liu et al. (2023) achieve strong performance through spatial-phase analysis but remain unimodal and do not quantify uncertainty. Multi-modal detectors developed in 2022–2024, including audio-visual transformers and CLIP-based detectors, demonstrate the value of combining modalities but typically rely on deterministic fusion. The proposed framework brings together three complementary modalities and adds an explicit probabilistic fusion layer, achieving accuracy comparable to or exceeding state-of-the-art results on FaceForensics++ while providing calibrated confidence estimates that are not available in most competing systems. Furthermore, its lower false-positive rate—6.5 percent compared with 9.2 percent for MesoNet—is particularly meaningful in operational scenarios such as journalism and law enforcement, where erroneous accusations of manipulation can cause significant harm.

Strengths of the Proposed Framework

The framework exhibits two principal strengths. The first is improved generalisation: by combining visual, temporal and emotional cues, the system reduces its reliance on any single type of artefact and is therefore less vulnerable to the dataset biases highlighted by Dolhansky et al. (2020). The second strength is uncertainty awareness. Because the Bayesian fusion module produces a posterior distribution rather than a point estimate, the system can flag low-confidence predictions for human review and is more robust against mild adversarial perturbations that often fool over-confident detectors. Together, these properties make the framework more suitable for real-world deployment than purely deterministic alternatives.

Practical Implications

The practical implications of this work extend across several domains. In media forensics and journalism, the system can be deployed as a browser plug-in or as a desktop verification tool that assists fact-checkers in evaluating user-submitted videos before publication, with the uncertainty score providing a transparent indicator of when human expertise should override the automated decision. In social-media moderation, the framework can be integrated into upload pipelines to triage high-risk content, prioritising clearly fake or clearly real items for automated handling and routing uncertain items to human moderators, thereby improving both throughput and accuracy. In financial services and identity verification, the same pipeline can be applied to video know-your-customer (KYC) checks to detect synthetic identities and replay attacks, where the

explicit confidence score is particularly valuable for regulatory auditability. In law enforcement and the judicial system, the framework can serve as an initial screening tool for digital evidence, with its calibrated uncertainty providing a more defensible basis for expert testimony than systems that produce only a binary verdict. Finally, in the educational and policy spheres, the modular nature of the framework makes it a useful platform for teaching students and policymakers about the strengths and limitations of current deepfake detection technology.

6. CONCLUSION

This study has presented a multi-modal deepfake detection framework that combines an Xception spatial stream, a DeepFace emotion stream and a SlowFast temporal stream through a Bayesian fusion module implemented in Pyro. The system was evaluated on FaceForensics++ using five-fold cross-validation and achieved 92.3% accuracy and an F1-score of 0.92, outperforming the Xception baseline by 5.1 percentage points in accuracy and 5.7 percent in F1-score, while reducing the false positive rate from 9.2% to 6.5%. The ablation study demonstrated that each modality contributes complementary signal and that the Bayesian fusion module provides a small but statistically significant additional improvement, accompanied by calibrated confidence scores that are valuable for downstream decision-making. The framework operates at 40 frames per second on GPU and 8.3 frames per second on CPU, making it suitable for a range of real-time and near-real-time applications in media forensics, social-media moderation, identity verification and law enforcement. The principal contributions of the paper are therefore a unified multi-modal architecture, an uncertainty-aware Bayesian fusion layer and a comprehensive empirical evaluation that demonstrates both accuracy and practical viability. Future work will focus on lightweight temporal models, cross-domain generalisation, audio-visual fusion, adversarial robustness and richer explainability mechanisms, with the broader goal of providing trustworthy tools for detecting synthetic media in an environment where generative models continue to advance at a rapid pace.

7.LIMITATION AND FUTURE SCOPE OF STUDY

Limitations of the Study

Notwithstanding these contributions, the study has several limitations that should be acknowledged. The first limitation concerns computational cost: the SlowFast 3D convolutional network introduces roughly 120 milliseconds of latency per frame on CPU, which restricts the applicability of the full pipeline in resource-constrained settings and on mobile devices. The second limitation is dataset coverage. Although FaceForensics++ remains a widely used and challenging benchmark, it does not fully reflect the variety of compression artefacts, lighting conditions and demographic distributions present in real-world social-media videos, and performance may degrade when the system is applied to cross-domain data such as the Deepfake Detection Challenge dataset or WildDeepfake.

The third limitation is that the present version of the system focuses on the visual modality and does not yet incorporate audio cues, which are increasingly important for detecting modern audio-visual deepfakes. The fourth limitation is that the Bayesian fusion module assumes a relatively simple variational family (Beta-distributed weights), and richer posterior approximations may yield further gains in calibration. Finally, while the false-positive rate has been reduced, no detector can be expected to be fully robust against adaptive adversaries who train generators specifically to evade the deployed system.

Future Scope of the Study

Several directions for future work follow naturally from the limitations identified above. A first priority is the development of lightweight architectures for the temporal stream, for example by replacing SlowFast with a combination of MobileNetV3 and LSTM units or with efficient transformer variants that can run at video frame-rate on commodity CPUs and mobile devices. A second direction is to evaluate and adapt the framework on diverse cross-domain datasets, including the Deepfake Detection Challenge, Celeb-DF and WildDeepfake, possibly using domain-adversarial or self-supervised pre-training to improve generalisation. A third direction is the integration of audio analysis and audio-visual synchronisation features, so that the system can detect manipulations that affect only the voice track or only the lip movements. A fourth direction concerns adversarial robustness, including the integration of gradient masking (Singh, 2020), certified defences and adversarial training to make the detector resilient to evasion attacks. A fifth direction is to enhance explainability by adding attention maps and Shapley-style attributions that clarify which modality and which spatial region contributed most to a given decision, which is essential for forensic and judicial use cases. Finally, future work should explore richer Bayesian formulations, including hierarchical priors and full posterior sampling, in order to obtain even better calibrated uncertainty estimates.

8. CONFLICT OF INTEREST

The authors declare that there are no known financial or non-financial conflicts of interest associated with this manuscript.

9. FUNDING ACKNOWLEDGEMENT

The authors received no financial support for the research, authorship, and/or publication of this article.

10. AI DISCLOSURE STATEMENT

AI-assisted tools were used solely to improve the manuscript's language quality, clarity, and presentation. Grammarly and QuillBot were utilized for assistance with grammar, style, and paraphrasing. Turnitin was employed to check plagiarism and potential AI-

generated content. No AI tool was used to generate original scholarly content, data, analyses, or conclusions. The authors take full responsibility for the originality, accuracy, and integrity of the manuscript.

11. REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE.DOI: 10.1109/WIFS.2018.8630761

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The Deepfake Detection Challenge dataset. arXiv preprint arXiv:2006.07397.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning (ICML), 1050–1059.

Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

Heo, Y., Choi, Y., Lee, Y., & Kim, H. (2023). Deepfake detection via combining frequency and spatial features with vision transformer. Sensors, 23(7), 3489.

Khan, S. A., & Dai, H. (2021, October). Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 1821-1828).

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., ... & Zhao, J. (2020, October). Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4318-4327).

Liu, Y., Zhu, X., Zhao, X., & Cao, Y. (2023). Spatial-phase shallow learning for deepfake detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 45(3), 2793–2807.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1–41.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, *3*(1), 80-87.

Singh, A. (2020, March 15). How generative adversarial networks (GAN) work. DataDrivenInvestor.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024, March). Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 5, pp. 5052-5060).

Wightman, R. (2019). PyTorch image models (timm) [Computer software]. GitHub repository.

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., & Wu, B. (2023). Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*.

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15023-15033).

AUTHOR'S BIO

Mr. Ashutosh Bhushan: - Mr. Ashutosh Bhushan is AI engineer, researcher, and communicator studying in Monash University, Caulfield campus. His research interests are AI ethics, Deep learning, Communication design and Social impact Application of AI.

AUTHOR CONTRIBUTIONS

Mr. Ashutosh Bhushan: Conceptualized the study, wrote the original draft, contributed to reviewing and editing the manuscript, and editing the manuscript.