Prototype for Deepfake Detection and Content Authenticity for Small Digital Media Creators

Susmit Acharya[1], Soma Chakraborty (advisor)[2]
[1,2]Department of Computer Science, Auxilium Convent School, Barasat, Kolkata, India
doi.org/10.64643/IJIRTV12I6-186359-459

Abstract – The rapid evolution of deepfake technology has made synthetic image and video manipulation accessible to the general public, threatening the credibility of digital content and the livelihoods of small media creators. Existing deepfake detection systems, while effective in controlled or enterprise contexts, remain computationally intensive, and inaccessible to independent creators who lack high-performance infrastructure. This research investigates and develops a lightweight, AI-powered authenticity verification framework specifically designed for small digital content creators such as youtubers, educators, and independent filmmakers. By integrating techniques like spatial, temporal, and multimodal consistency analysis with resource-efficient architectures like MobileNet V3, the study aims to achieve real-time detection on consumer-grade hardware. The proposed system leverages open-source deepfake datasets such as FaceForensics++ and Celeb-DF to train and validate detection robustness under common real-world conditions such as compression, low resolution, and varied lighting. Furthermore, the framework introduces a user-centric interface that enables authenticity scoring. The experimental framework emphasizes a balance between accuracy, latency, and usability, thereby bridging the gap between academic deepfake forensic studies and the practical needs of independent creators. The outcome is a deployable framework and operational roadmap that democratizes content authenticity verification, contributing to digital trust and ethical media creation.

Index Terms – Deepfake Detection, Content Authenticity, AI Forensics, Lightweight Neural Networks, Media Verification; Small Creators, Video Integrity, Algorithms, Machine Learning, Open-Source.

I. INTRODUCTION

1.1 Context of the Research

In today's digital media landscape, the rise of deepfake technology i.e. digitally generated synthetic media via artificial intelligence (AI) to alter visual or audio content has been a major hurdle to content authenticity and public safety. State-of-the-art generative models like Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2], and diffusion models [3] allow for creation of hyper-realistic manipulated image and video clips, and are being unethically used to fabricate media and spread misinformation throughout the internet, thus harming digital creators. Suwajanakorn et al.'s "Synthesizing Obama" [4], for instance, showed realistic audio-driven facial animation, highlighting the ease with which deepfakes can and continue to proliferate the internet with fake content. While such technologies have artistic uses, they have also been used to facilitate widescale spread of misinformation, creation of non-consensual images, fraud, and political manipulation.

For decentralized small digital content producers like youtubers, teachers, reporters, etc. the rise in artificial media is a two-fold menace with possible misrepresentation of their likeness and the loss of audience trust. Unlike big media companies, such creators do not have enterprise-grade content verification or forensic capabilities. Current deepfake detection methods like Intel's FakeCatcher, Sensity AI, etc. while effective, usually require high-end hardware (GPUs, special compute nodes) and expertise, rendering them incompatible for most decentralized, resource-constrained users. Accordingly, there is a strong need for an inexpensive, and easy-to-use verification system that protects the content integrity that is generated by creators yet is still accessible.

1.2 The Problem

Deepfake detection techniques have made great progress, but real-world implementation is still mostly feasible within organizational and corporate premises. Detection methodologies are predominantly computationally demanding, involving deep

convolutional networks [5] (such as XceptionNet [6], EfficientNet [7]) or multimodal transformer structures [8] that cannot practically use personal devices. Their applications, thus, remain restricted for small content producers who usually use regular laptops or inexpensive cloud platforms. Lack of authenticity tools that serve creators, poses the risk of manipulated content in mainstream content, and if they remain unchecked and their credibility gets diluted, independent producers remain exposed to face-impersonation or slander.

1.3 Background and Related Studies

Recent scholarly advances in deepfake detection have named multiple methodologies for deepfake detection. Frame-level spatial methods [9] identify pixel-level anomalies and blending seams with compact convolutional models like MobileNet V3 [10] , while temporal and multimodal methods [11] use sequence learning and audio-visual synchronization for added robustness. Standard public deepfake datasets like FaceForensics++ and Celeb-DF have allowed for regularized benchmark training and tests in research studies.

Even with these improvements, high-performing models favor accuracy at the cost of accessibility. They are tuned for GPU compute, initially trained at high resolution, and do not usually consider compression or low-light conditions characteristic of web image and video. Yan Wang et. al.'s 2024 study on Multi-Domain Awareness for Compressed Deepfake Videos Detection [12] shows that when the mainstream detection models are subjected to domain shift i.e. when they are assessed with manipulated methods they haven't seen or low-quality video, and the results show that the accuracy of the mainstream deepfake detection plummets in such cases. Thus, in spite of adequate foundations for deepfake detection in theory, resource-efficient adaptation is, again and again, a persistent thread throughout the available literature on deepfake detection.

1.4 Identified Research Gap

Existing deepfake detection frameworks largely overlook the needs of small, resource-constrained creators. Most available models fail to generalize beyond laboratory datasets and require hardware or expertise that independent users cannot feasibly maintain. No open-source, lightweight system currently bridges advanced detection algorithms with user-friendly interfaces suited to non-technical creators. This research directly addresses that gap by focusing on accessibility and computational efficiency.

1.3 Objectives of the Research

1. To develop a lightweight, AI-driven authenticity verification framework optimized for small digital media creators.
2. To integrate spatial, temporal, and multimodal deepfake detection methods into a resource-efficient model architecture (e.g. MobileNet V3).
3. To evaluate detection accuracy, latency, and robustness under real-world conditions including compression, noise, and variable lighting.
4. To design an intuitive interface producing authenticity scores.

1.5 Scope and Limitations

This research is confined to the domain of image-based deepfake detection and content authenticity verification, emphasizing usability for small creators. The system focuses on detecting manipulations such as facial identity swaps, and expression transfers within short-form digital content. Due to computational and data-access constraints, the study utilizes publicly available datasets (FaceForensics++ and Celeb-DF). While the proposed model aims for efficiency on consumer-grade hardware, its performance may vary across unseen generative models or extreme compression levels. The study also does not extend to proactive prevention or blockchain-based content traceability. The research only provides a foundational verification layer that can complement future integrity-preserving frameworks.

II. METHODOLOGY

2.1 The Approach

This research adopts a prototype-driven, empirical approach aimed at creating an effective deepfake detection and authenticity verification system for small digital content producers. The general methodology incorporates both theoretical and applied facets of AI-driven digital forensics. This starts with dataset construction from open-source, which is then

followed by the construction of a lightweight neural network designed for consumer-grade hardware. The approach focuses on three major priorities namely efficiency, precision, and usability making sure that the resulting model has high detection accuracy without relying on massive-scale computational resources. After training and verification, the system will be deployed as a web application.

2.2 Dataset Preparation

A mix of available deepfake datasets released into the public domain constitutes the dataset of this study. Baseline training and benchmarking are carried out using standardized datasets, which cover a large range of forgery methods, compression amounts, and image quality.

Each clip falls into one of two categories: genuine or manipulated. The manipulated clips are created with open-source libraries, which produce photorealistic fake copies of original clips. For enhanced accuracy of the model, separation is made between validation sets and training sets, so that the model does not tend to overfit on individual faces. The training dataset consists of 13000+ images, and the validation dataset consists of 3000+ images, for evaluation of the model's learning and accuracy. Preprocessing consists of facial recognition, alignment, and resizing of images for normalization on input. This guarantees both fake and real samples have similar dimensional and visual features which is suitable for training on a neural network.

2.3 Creating the model

The proposed architecture prioritizes computational efficiency and deployment flexibility. Two primary model strategies are explored:

1. Custom Lightweight CNN: A small convolutional neural network targeting local anomaly detection. This structure uses small convolution kernels and shallow depth for micro-level anomalies like blending seams, unnatural-looking edges, or texture aberrations. Following the footprint of MobileNet V3, this setup strives for close–to–real–time inference efficiency.
2. Transfer Learning from Pretrained Models: We fine-tune pretrained lightweight models like MobileNetV3. Through freezing convolutional earlier layers and fine-tuning top layers on forgery databases, this method declines training expense greatly and facilitates deployment on standard CPUs or GPUs.

Training is done in Python, taking advantage of libraries such as TensorFlow and PyTorch for model building, OpenCV for image processing, and NumPy and Pandas for data manipulation. Measuring metrics are precision, recall, F1-score, and latency, to balance accuracy and responsiveness. Cross-dataset validation (e.g., training on FaceForensics++ and testing on Celeb-DF) is used to test model generalization across unseen fake-generation approaches. A multi-step face detector, based on the "waterfall" pipeline of Balafrej, et al. (2024) [13], integrates fast detectors with fallback detectors for enhanced robustness.

III. RESULTS AND DISCUSSIONS

3.1 Evaluation Metrics

To assess the performance and practicality of the proposed deepfake detection model, we define both quantitative and qualitative metrics.

Quantitative Metrics:

- Detection Accuracy: Measures the overall correctness of classification between authentic and manipulated content.
- Precision and Recall: Evaluate how effectively the model identifies fake content while minimizing misclassification of authentic samples.
- F1-Score: Represents the balance between precision and recall, indicating the model's overall reliability.
- Inference Latency: Captures the average time taken to process and evaluate a video, ensuring near real-time performance.

Qualitative Metrics:

- User Experience (UI/UX): Evaluates how intuitively users can upload content, view results, and interpret authenticity outputs.
- Interpretability: Measures the clarity of visual outputs with confidence percentages, helping users easily understand whether their content is authentic or manipulated.

3.2 Model Accuracy and Performance Analysis

Table 1: *Evaluation Summary of MobileNetV3-based Deepfake Detector*

Metric	Value
Accuracy	0.6804
Precision	0.6752
Recall	0.9950
F1 Score	0.8045
ROC-AUC	0.7008
Inference Time per Frame	0.00656 s (≈152 FPS)

The model's overall accuracy is 68.04%, and it has a high recall of 99.5%, meaning that it identifies almost all the manipulated material correctly. However, its precision is 67.5%, which means that there were some true videos wrongly identified as fake, consistent with the confusion matrix finding that "authentic" samples had a low true positive rate of 7%. This trade-off indicates that the model is sacrificing specificity (not flagging false positives) for sensitivity (flagging all fakes), which is a good property in forensic screening applications. The ROC-AUC value of 0.70 also validates moderate discriminative power between true and spoofed samples. In addition, an inference rate of 152 FPS proves compatibility with real-time or near real-time applications.

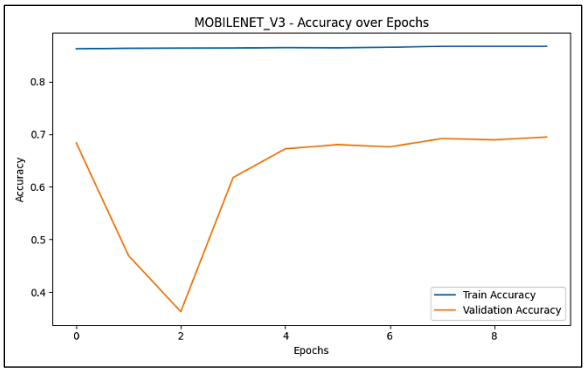
Table 2: Class-wise Performance

Class	Precision	Recall	F1-Score	Support
Authentic	0.87	0.07	0.13	816
Fake	0.68	0.99	0.80	1590

Although the model is extremely sensitive to false content (recall ≈99%), genuine samples are under-detected. The imbalance is indicative of data bias or overfitting to false features — to be improved in future training by class rebalancing or loss weighting.

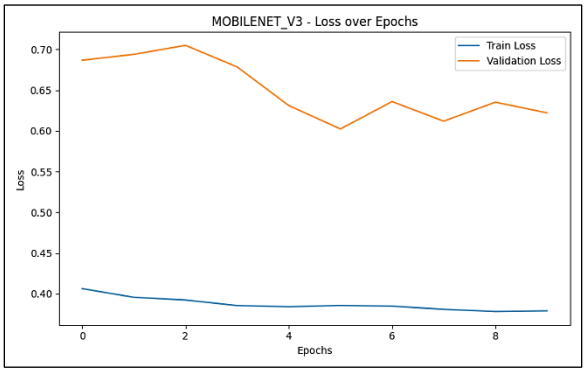
3.3 Visual Outputs

Figure 1: *Accuracy progression for MobileNet V3 across epochs*



This visualization summarizes the internal performance trend of the model, interpreting how well it differentiates between real and synthetic faces. It can be noticed that the training accuracy is steadily high at ≈ 0.87–0.88, while the validation accuracy sees quite a variation before converging to about 0.70. The divergence in performance measures proves partial overfitting of the model-which learned the patterns of training data well but does not generalize as well to unseen samples. Visualization like this lets the developers interpret learning stability, diagnose the imbalance of real and fake data distributions, and spot the need for data augmentation or fine-tuning. Hence, these performance curves enhance the transparency and traceability of the tool by enabling end-users and evaluators to assess the reliability of its predictions.

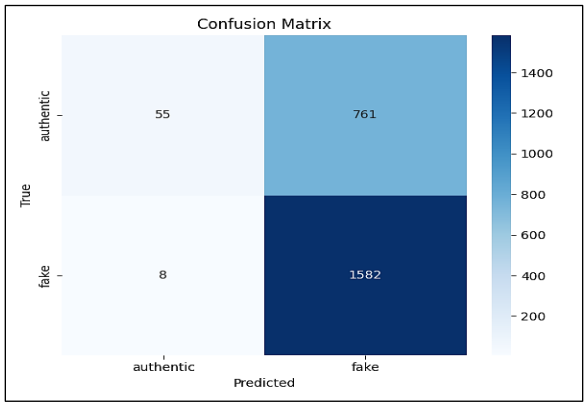
Figure 2: *Training and validation loss for MobileNet V3 over epochs*



The loss curve presents steady behavior for the training loss, converging to the value of approximately 0.38, which means effective optimization. However, the validation loss fluctuates for a while before gradually going down. This could indicate that the model keeps learning generalizable features, rather than overfitting early. The gap between the two curves is pretty consistent, which means a moderate level of

bias that may be due to dataset imbalance or high intra-class variability common in facial forgery datasets. This behavior reflects the network's ability to minimize reconstruction and classification errors while sustaining good generalization properties on unseen validation data. Such trends demonstrate a relatively stable training process, implying that the selected architecture and learning rate were appropriate for the dataset scale and preprocessing pipeline.

Figure 3: *Confusion Matrix for the MobileNetV3 Model*

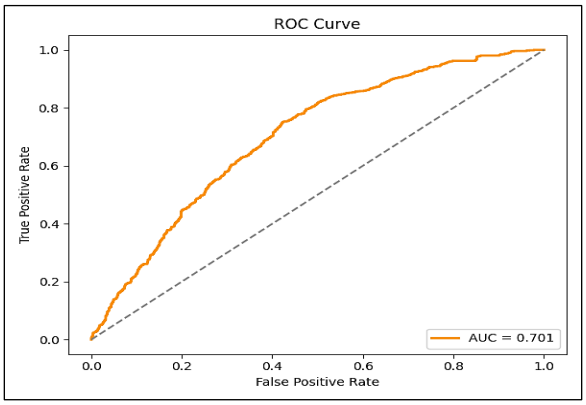


Through the matrix, it is seen that the model accurately picked 1,582 of the fake samples but misclassified 8 of the fake samples as real. On the other hand, it accurately picked only 55 real samples, misclassifying 761 real frames as fake. This is an indication that although the model is very sensitive to picking manipulated content (low false negatives), it has low specificity towards real content (high false positives). This pattern suggests an overfitting prejudice towards the "fake" class, likely caused by dataset imbalance, feature dominance of the tampered textures, or lack of proper representation of authentic samples within the training set. Despite the general accuracy being high because of dominance of the fake class, real-world applicability needs to be improved upon when separating authentic instances.

Future training iterations can address this by:

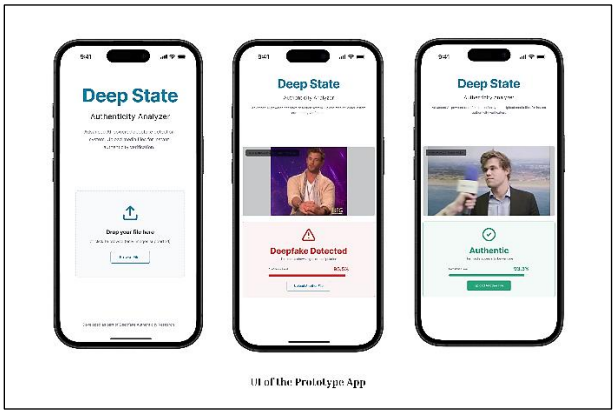
- Introducing class weighting or oversampling authentic frames.
- Employing focal loss to penalize dominant class bias.
- Enhancing the dataset with diverse real samples across lighting and compression conditions.

Figure 4: *ROC Curve for the MobileNet-V3 model*



ROC (Receiver Operating Characteristic) curve depicts the True Positive Rate (sensitivity) versus the False Positive Rate (1 – specificity) for different classification thresholds. Random classification is indicated by the diagonal grey line (AUC = 0.5), while the orange curve depicts the discrimination capability of the model between real and fake samples. In this experiment, the AUC is 0.701, meaning that it has a moderate discrimination capability. That means while it may generate deepfakes much better than random guessing, its prediction will have noticeable overlap between genuine and manipulated data distributions. This behavior is in line with the imbalance observed in the confusion matrix (Figure 8), where dominating fake detections are probably due either to skew in the dataset or overfitting towards synthetic samples. These results collectively point to the fact that MobileNetV3, though lightweight and computationally efficient, may require further fine-tuning, regularization, or balanced data augmentation to achieve stronger generalization performance on unseen deepfake samples.

Figure 5: *App Interface for the End-User Application*



The user interface (UI) of the deepfake detector is crafted around simplicity and ease of use. As displayed, the app offers a minimalist upload interface where the user can upload an image or a short video clip to be tested for authenticity. When uploaded, the content is processed by the system and the outcome is displayed in an easy-to-read manner with a clear binary answer ("Authentic" or "Deepfake") with the respective confidence percentage. Such a simple feedback mechanism makes it so that even novice users can easily understand the results without confusion. The layout of the interface prioritizes usability and trust in accordance with human-centered design. Decisive typography, consistent color cues, and adaptive visual feedback facilitate easy differentiation of results and comprehension of their implications. Once fully deployed, user feedback can further refine these design elements to enhance comprehension, efficiency, and overall experience. Thus, it can provide creators with both verification confidence and explainability of the results generated.

IV.CONCLUSION

This work introduces a feasible system for deepfake detection and content authenticity verification for small digital media producers. Through the incorporation of lightweight neural network designs, open-source datasets, and low-resource design techniques, the system bridges the gap between research forensics and actual usability. The model's compatibility with consumer-grade hardware positions the system as an accessible low-cost remedy for ensuring content integrity in a fast-changing digital world. Though dataset dependency and dynamically changing manipulation methods remain as limitations, the work lays a foundation for democratized fact-checking of content. This study adds to the advancements in increasing dataset diversity, enhancing cross-model generalization, and fusing proactive authenticity preservation, thus, aiding the robustness of digital trust, empowering independent creators, and maintaining the credibility of online visual media.

ACKNOWLEDGEMENT

I express special gratitude to Mrs. Soma Chakraborty, my Computer Science instructor, for her direction, encouragement, and guidance throughout this research. Her encouragement has gone a long way towards shaping both the technical and analytical aspect of this work. I also appreciate the constant encouragement of peers and well-wishes who encouraged me to embark on this project with dedication and perseverance.

REFERENCE

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio. (2014). Generative Adversarial Networks. https://arxiv.org/pdf/1406.2661

[2] Kingma, D., & Welling, M. (2019). An Introduction to Variational Autoencoders. Foundations and Trends® in Machine Learning, 12(4), 307–392. https://arxiv.org/pdf/1906.02691

[3] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, & Ming-Hsuan Yang. (2025). Diffusion Models: A Comprehensive Survey of Methods and Applications https://arxiv.org/pdf/2209.00796

[4] Suwajanakorn, S., Seitz, S., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph., 36(4). https://doi.org/10.1145/3072959.3073640

[5] Mallat Stéphane 2016Understanding deep convolutional networksPhil. Trans. R. Soc. A.37420150203 https://doi.org/10.1098/rsta.2015.0203

[6] V, A., & Joy, P. (2023). Deepfake Detection Using XceptionNet. In 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) (pp. 1-5). https://doi.org/10.1109/RASSE60029.2023.10363477

[7] Mingxing Tan, & Quoc V. Le. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. https://arxiv.org/pdf/1905.11946

[8] Peng Xu, Xiatian Zhu, & David A. Clifton. (2023). Multimodal Learning with Transformers: A Survey. https://arxiv.org/pdf/2206.06488

[9] Li, M., Zhang, X.P., & Zhao, L. (2025). Frame-Level Temporal Difference Learning for Partial Deepfake Speech Detection. IEEE Signal Processing Letters, 32, 3052–3056. https://arxiv.org/pdf/2507.15101v1

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, & Hartwig Adam. (2019). Searching for MobileNetV3. https://arxiv.org/pdf/1905.02244

[11] Valentin Vielzeuf, Stéphane Pateux, & Frédéric Jurie. (2017). Temporal Multimodal Fusion for Video Emotion Classification in the Wild. https://arxiv.org/pdf/1709.07200

[12] Yan Wang, Qindong Sun, Dongzhu Rong, & Rong Geng (2024). Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts. Computer Vision and Image Understanding, 247, 104072. https://doi.org/10.1016/j.cviu.2024.104072

[13] Balafrej, I., Dahmane, M. Enhancing practicality and efficiency of deepfake detection. Sci Rep 14, 31084 (2024). https://doi.org/10.1038/s41598-024-82223-y

Additional Resources Used

Dictionaries:

1. Cambridge Dictionary. Retrieved from: https://dictionary.cambridge.org/us/dictionary/english/

GitHub Repositories:

1. FaceForensics++. Retrieved from: https://github.com/ondyari/FaceForensics
2. Celeb-DF. Retrieved from: https://github.com/yuezunli/celeb-deepfakeforensics
3. Developed Prototype: https://github.com/SusmitAcharya/deepfake_detector