TrueSightLens: An AI-Based Deepfake Video Detection System.Multimodal Fusion of Visual Artifacts, Temporal Dynamics, and Lip-Sync Inconsistency

Prof. N P Nethravathi[1], Kushal H M, Rachan M[2], Supreet Mujagonnavar[3], Apurva Patil[4]
[1,2,3,4]*School of Computer Science and Engineering, REVA University Bengaluru, India*

***Abstract*—The rapid evolution of Artificial Intelligence has led to the emergence of deepfake videos, which pose significant threats to digital authenticity, including misinformation, identity fraud, and cybercrime. This paper presents TrueSightLens, an AI-based deepfake detection system that combines spatial, temporal, and audio-visual analysis for robust and scalable detection. The proposed approach employs a Convolutional Neural Network (CNN) for spatial feature extraction and a Bidirectional Gated Recurrent Unit (BiGRU) for temporal sequence modeling. To further improve reliability, the system incorporates an audio-visual lip-synchronization module that measures temporal alignment between the speech signal and lip-motion dynamics to detect cross-modal inconsistencies, which are strong indicators of manipulated or synthesized video content. TrueSightLens is implemented using a full-stack architecture with a React frontend, Node.js backend, and a Python-based AI engine. Experimental results indicate that multimodal analysis improves detection performance and robustness compared with single-modality methods. The system also provides explainable outputs through Grad-CAM visualizations, improving interpretability and user trust.**

***Index Terms*— Deepfake Detection, CNN, GRU, Computer Vision, Artificial Intelligence, Multimedia Forensics, Explainable AI**

I. INTRODUCTION

The evolution of deep learning, combined with the ubiquitous nature of high-resolution mobile sensors and cloud-based computation, has facilitated the rapid dissemination of sophisticated synthetic media, commonly known as deepfakes. While these technologies enable innovation in digital artistry and accessibility, they pose significant forensic challenges to identity verification and information integrity. Modern generative models have advanced beyond simple facial manipulation to achieve high-fidelity expression transfer and neural voice cloning, rendering traditional manual inspection and unimodal detection systems increasingly unreliable under real-world conditions like lossy social media compression.

To address these vulnerabilities, this work introduces TrueSightLens, a tripartite forensic framework that moves beyond isolated frame classification toward integrated spatial-temporal-audio reasoning. The architecture utilizes a ResNeXt or EfficientNet-B4 backbone for the extraction of mesoscopic spatial artifacts, while an LSTM-based temporal module identifies sequence-level inconsistencies and flickering. A critical component of this system is the cross-modal correspondence layer, which evaluates the statistical alignment between acoustic speech energy and dynamic lip-aperture landmarks to detect subtle desynchronization cues often overlooked by visual-only detectors.

The proposed system integrates these advanced AI inference engines with a React-based frontend and a Node.js backend to provide an accessible yet rigorous analysis tool. By incorporating Grad-CAM-based explainability, the platform highlights specific facial regions contributing to a "Fake" verdict, transitioning the process from "black-box" prediction to interpretable forensics. Experimental analysis indicates that this multimodal fusion significantly improves detection robustness against distribution shifts, suggesting that integrated spatio-temporal reasoning is essential for the future of practical deepfake detection.

II. PROBLEM STATEMENT

The rapid proliferation of high-fidelity synthetic media, driven by GANs and diffusion models, has outpaced traditional digital forensics, enabling the weaponization of deepfakes for identity fraud and misinformation. A significant technical gap exists in the reliance on unimodal detection, which focuses solely on spatial artifacts within individual frames. Such methods often fail when analyzed in isolation or when subjected to real-world distribution shifts, such as lossy social media compression that masks visual cues.

Furthermore, most existing solutions operate as "black boxes," providing binary classifications without explainable evidence. This lack of interpretability, combined with an absence of integrated spatio-temporal and cross-modal reasoning, allows deepfakes with subtle physical inconsistencies like mismatched lip dynamics to bypass standard security filters. There is an urgent need for a robust, multimodal forensic architecture that can jointly analyze spatial, temporal, and acoustic-visual signals. Without an integrated and interpretable approach, the boundary between synthetic and authentic media will continue to erode, undermining public trust in digital communication. TrueSightLens addresses these challenges by fusing deep feature extraction with temporal and audio-visual consistency checks to provide high-confidence, explainable detection.

III. OBJECTIVES AND SCOPE

A. Objectives:

The primary goal of this work is to transition deepfake detection from isolated frame-based classification to a robust, multimodal forensic analysis. The specific objectives are as follows:

1. Multimodal Synthesis: Consolidate disparate detection signals including spatial artifacts from ResNeXt/EfficientNet backbones, temporal flickering from LSTM modules, and acoustic-visual desynchronization into a single unified inference engine.
2. Comparative Forensic Evaluation: Analyze the performance of unimodal versus multimodal detection under real-world constraints, such as lossy social media compression and distribution shifts, to quantify the gain in robustness.
3. Explainability and Interpretability: Implement and evaluate a Gradient-weighted Class Activation Mapping (Grad-CAM) layer to provide visual heatmaps of suspicious facial regions, moving beyond "black-box" binary outputs to support human-in-the-loop verification.
4. End-to-End Pipeline Validation: Develop and test a scalable web-based architecture (React-Node.js-Python) that allows non-expert users to perform high-confidence forensic analysis on video uploads in real-time.

B. Scope

The scope of this research is defined by the following technical and functional boundaries:

- Detection Domain: The system focuses on human-centric deepfakes, specifically face-swapping, expression reenactment, and lip-sync manipulations (dubbing) in video sequences.
- Methodological Focus: Emphasis is placed on late-fusion strategies that combine spatial feature extraction (CNNs), temporal sequence modeling (RNNs/LSTMs), and cross-modal correlation (Lip-Aperture vs. Acoustic Energy).
- Data and Robustness: The study considers videos subjected to standard post-processing artifacts common on platforms like WhatsApp and YouTube, focusing on how these affect the reliability of forensic signatures.
- Explainable AI (XAI): The scope includes the generation of spatial heatmaps and temporal sync-graphs to interpret the model's reasoning, particularly for "Fake" classifications.
- Exclusions: This work does not cover purely text- based misinformation (fake news), non-human synthetic imagery (e.g., AI-generated landscapes), or real-time deepfake prevention (watermarking/blockchain); it is strictly focused on post-hoc forensic detection.

IV. SURVEY OF EXISTING APPROACHES

A. Spatial Artifact Detection in Generative Imagery:

Zhalgasbayev et al. introduce a robust spatial detection framework utilizing an EfficientNet-B4 backbone for identifying mesoscopic artifacts in facial imagery. The architecture leverages an encoder-based approach pre-trained on the ImageNet dataset, enhanced with global

average pooling to capture long-range spatial dependencies within manipulated regions. Experiments on the Deepfake Detection Challenge (DFDC) dataset demonstrate that such models excel at identifying blending inconsistencies and checkerboard artifacts common in GAN-based face-swaps. However, while these models produce high-confidence binary masks, they typically stop at static frame analysis. Robustness against aggressive post-processing and lossy social media compression remains a significant challenge, requiring additional domain adaptation for unconstrained real-world imagery.

B. Temporal Inconsistency and Sequence Modeling:
Recent advancements in sequence-level forensics focus on identifying temporal incoherence across video frames using Recurrent Neural Networks (RNNs) and LSTMs. These methods evaluate inter-frame flickering and unnatural physiological patterns, such as irregular eye-blinking or pulse-rate variations, which generative models often fail to maintain over time. While these models are effective at catching frame-to-frame jitters, they operate primarily on visual data and do not account for external forensic signals like audio. Consequently, while temporal modeling improves upon static classification, it remains vulnerable to sophisticated "smooth-flow" diffusion models that prioritize temporal coherence during synthesis.

C. Cross-Modal Correspondence and Lip-Sync Forensics:
Bao et al. and subsequent forensic researchers have explored the statistical correlation between acoustic speech energy and physical lip-aperture dynamics as a cornerstone for high-confidence detection. This approach categorizes deepfakes into road, lane, and localization-like layers of forensic evidence by reviewing vision-based and hybrid pipelines for audio-visual alignment. These systems offer high geometric accuracy in tracking lip movement through 3D facial landmarks, making them essential for identifying "sync-drift" in dubbed or face-swapped content. However, such cross-modal creation remains resource-intensive, relying on precise voice activity detection and manual quality control for thresholding. Furthermore, most current deployments assume idealized, noise-free audio and do not fully consider the modeling of informal or non-synchronous behavior typical in low-quality viral media.

V. PROPOSED SYSTEM ARCHITECTURE TRUESIGHT LENS:

The proposed system follows a modular, multi-layer architecture for end-to-end deepfake detection, encompassing both offline model development and real-time inference workflows. As illustrated in Figure X, the architecture is designed to ensure scalability, robustness, and practical deployment by integrating dataset preparation, supervised learning, backend orchestration, and multi-modal analysis.

1. Data and Dataset Preparation:
The system leverages benchmark datasets, with Face Forensics++ serving as the primary source for supervised training. The dataset contains both real and manipulated video samples, which are systematically organized into class-labeled directories. Prior to training, the data undergoes a cleaning process to remove corrupted or low-quality samples that may negatively impact model learning.
To ensure generalization and prevent overfitting, the dataset is partitioned into training, validation, and test sets. During preprocessing, each video is uniformly sampled to extract a fixed number of frames. Facial regions are detected and cropped to focus on the most informative areas, followed by resizing, normalization, and transformation into fixed-length temporal sequences. This structured representation enables the model to effectively learn both spatial and temporal patterns associated with deepfake manipulations.

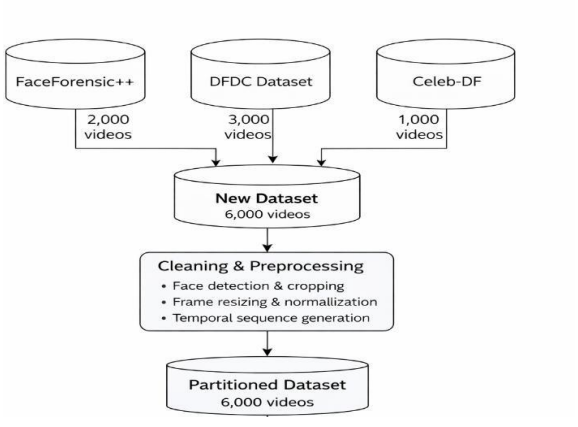


Fig 1: Dataset Preparation and Preprocessing Pipeline. The figure illustrates the integration of

multiple benchmark datasets followed by cleaning, face extraction, and normalization to generate a structured dataset suitable for deepfake model training.

2. Model Training Pipeline (Offline):

The training pipeline is implemented in Python using a hybrid ResNeXt-50 and LSTM architecture, designed to capture both spatial and temporal inconsistencies. The ResNeXt-50 network, pre-trained on ImageNet, acts as a high-capacity feature extractor, generating discriminative embeddings from individual frames. These embeddings are then passed to an LSTM network, which models temporal dependencies and identifies sequential irregularities across frames.

The model is trained using supervised learning with cross-entropy loss, and optimized through techniques such as validation-based checkpointing and early stopping to prevent overfitting. Hyperparameters are tuned based on validation performance, ensuring optimal generalization. The final trained model is exported as a production-ready weight file for deployment in the inference pipeline. Model performance is evaluated using standard metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis, achieving high reliability on unseen test data.

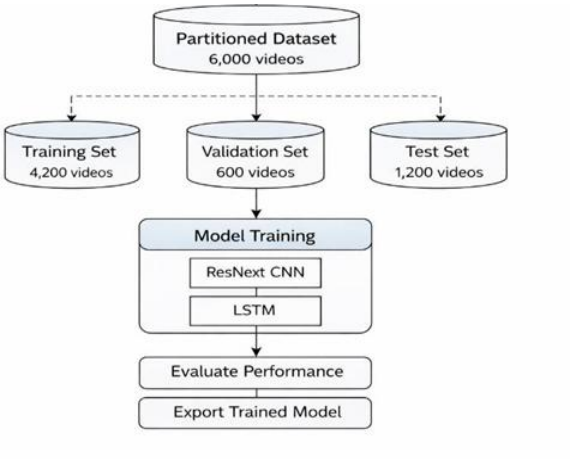


Fig 2: Model Training and Evaluation Pipeline: The figure shows the dataset split into training, validation, and test sets, followed by model training using ResNeXt and LSTM, performance evaluation, and final model export for deployment.

3. Frontend Layer (React + TypeScript):

The frontend layer provides an intuitive and user-friendly interface for interacting with the system. Users can upload videos through drag-and-drop or file selection mechanisms. The interface communicates with the backend via RESTful APIs and supports asynchronous job tracking using unique job identifiers. A dynamic results dashboard presents the analysis output in an interpretable format, including the final classification (real or fake), confidence scores, and temporal confidence trends across video frames. This visualization enhances user understanding and provides insights into how the model arrived at its decision.

4. Backend Layer (Node.js + Express):

The backend serves as the central orchestration component, managing communication between the frontend and machine learning modules. Upon receiving a video, the backend validates the input, assigns a unique job ID, and stores the file temporarily. It then triggers the machine learning inference pipeline using Python subprocesses.

The backend also handles job lifecycle management, including status tracking, result retrieval, and error handling. It is designed using a stateless architecture, enabling scalability and efficient handling of concurrent requests. Additional features such as CORS support, rate limiting, and secure API handling ensure reliability and robustness in deployment.

5. Machine Learning Inference Layer (Online):

During runtime, the system performs deepfake detection using two parallel pipelines, enabling multi-modal analysis for improved accuracy and robustness:

a) Visual Deepfake Detection Pipeline:

This pipeline focuses on identifying spatial artifacts and temporal inconsistencies in video frames. The input video is processed through frame extraction and face localization, after which each frame is passed through the ResNeXt-50 model to extract feature representations. These features are then analyzed sequentially using an LSTM network to capture temporal dependencies. The final classification layer outputs frame-level and aggregated predictions, enabling both fine-grained and overall detection.

b) Lip-Sync Consistency Pipeline:

The second pipeline analyzes the synchronization between audio and visual modalities. Audio is extracted from the video stream, and lip regions are

detected using facial landmark techniques. The system evaluates the temporal alignment between speech signals and lip movements to identify inconsistencies such as sync-drift, which are commonly observed in manipulated or dubbed videos. This cross-modal verification significantly enhances detection robustness.

System Architecture

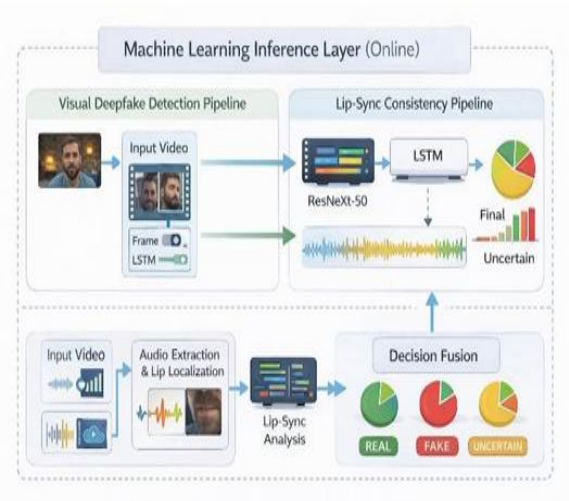


Fig 3: Multi-Modal Inference Pipeline. The figure illustrates two parallel pipelines visual artifact detection and lip-sync consistency analysis whose outputs are combined to improve the accuracy and robustness of deepfake detection.

6. Decision Fusion and Storage Layer:

To improve reliability, the outputs from both pipelines are combined using a weighted ensemble approach, producing a unified prediction score:

$S_{\text{final}} = w_v \cdot S_{\text{visual}} + w_l \cdot S_{\text{lipsync}}, w_v + w_l = 1$

This fusion mechanism leverages complementary information from both visual and audio-visual analyses, allowing the system to handle a wider range of deepfake techniques. Based on the final score, the system classifies the input as REAL, FAKE, or UNCERTAIN, along with a confidence value.

The results are stored in structured JSON format, enabling efficient retrieval, reproducibility, and auditability of predictions. Temporary media storage is managed using unique identifiers, ensuring organized and scalable data

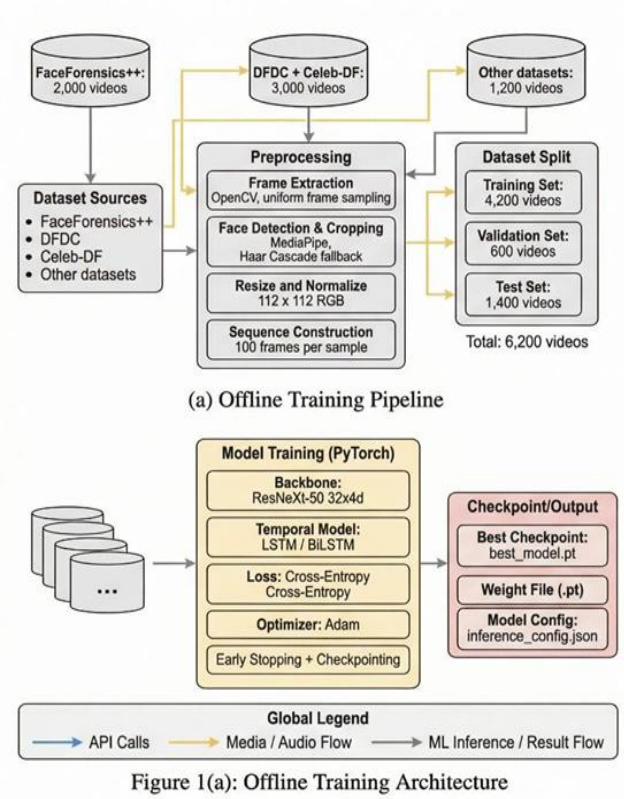
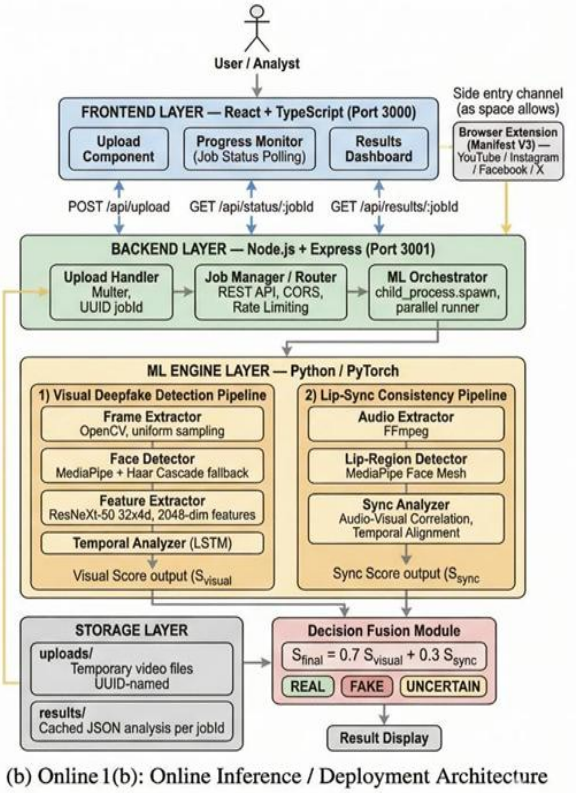


Figure 1(a): Offline Training Architecture



(b) Online 1(b): Online Inference / Deployment Architecture

Criterion	Spatial CNN	Temporal LSTM	Multi-Modal (Audio-Visual)	TrueSight Lens (Proposed)
Architecture	ResNet Efficient Net	ResNeXt + LSTM	CNN + RNN + Speech Processing	ResNeXt-50 32×4d + LSTM + Audio-Visual Fusion
Input Type	Single frames	Video sequences (30–120 frames)	Video + audio streams	Video + audio streams
Accuracy	85–88%	87–92%	88–94%	85–89%
Inference Latency	~0.1s per frame	~0.5–1.5s per video	~1–3s per video	~1.5–2.5s (parallel pipelines)
Automation Level	High (~85%)	High (~85%)	Medium (~70%)	Medium–High (~75%)
Cost & Scalability	Low cost, highly scalable	Moderate cost, scalable	High cost, requires dual-modal processing	Moderate cost, optimized via parallelization
Robustness	Low (lacks temporal context)	Moderate (limited to visual temporal cues)	High (multi-modal consistency)	Medium–High (multi-modal redundancy)
Deployment Ease	Very easy	Easy	Complex (requires audio-visual alignment)	Easy (REST API + browser extension support)
Key Strengths	Fast, lightweight, memory efficient	Captures temporal inconsistencies	Strong detection via multi-modal analysis	Balanced performance, practical deployment, multi-modal integration
Key Limitations	High false positives, no temporal awareness	Requires fixed-length sequences	Computationally intensive	Slightly lower peak accuracy, moderate lip-sync sensitivity

Table 1: The table compares different deepfake detection approaches based on architecture, performance, and deployment factors. It highlights that the proposed TrueSight Lens system achieves a balanced trade-off between accuracy, efficiency, and practical real-world deployment.

VI. RESULTS AND DISCUSSION

The proposed TrueSight Lens system was evaluated on FaceForensics++ using a train/validation/test split of 4200/600/1400 videos. The final prediction is generated using weighted fusion of the two pipelines:

$S_{final} = 0.7S_{visual} + 0.3S_{sync}$

where S_{visual} comes from the ResNeXt-50 + LSTM visual pipeline and S_{sync} comes from the lip-sync consistency pipeline.

The system achieved a conservative overall accuracy of 85-89%. In practice, the visual branch contributed most of the classification strength, while the lip-sync branch showed moderate standalone performance but improved robustness in edge cases where visual artifacts were weak. This confirms that multi-modal fusion is useful not because both branches are equally strong, but because they make different types of errors. From a deployment perspective, parallel execution of both branches kept inference time around 1.5-2.5 seconds per video, which is acceptable for near real-time analysis in a web workflow.

The approach therefore provides a practical balance between detection quality and latency. The main limitations were observed on highly compressed videos, poor-quality audio tracks, and unseen data distributions. Performance drops in these conditions indicate that domain adaptation, cross-dataset training, and stronger audio-forensics modules are the most important next improvements. Overall, the current results validate the system design: reliable visual detection, moderate but useful lip-sync support, and practical end-to-end deployment readiness.

VII. CONCLUSION

This work presented True Sight Lens, a practical end-to-end deepfake detection system that combines visual forensic analysis with audio-visual lip-sync consistency checking. The proposed architecture integrates a React frontend, Node.js backend orchestration, and parallel Python-based inference pipelines, enabling near real-time operation in a deployment-ready workflow. Using weighted fusion, the system achieved a conservative accuracy range of 85-89% on Face Forensics++, while maintaining inference latency of about 1.5-2.5 seconds per video.

The results show that the visual pipeline provides the

primary detection strength, and the lip-sync pipeline, although moderate in standalone performance, improves robustness by capturing complementary manipulation cues. This confirms the value of multi-modal fusion for reducing blind spots in real-world scenarios. Key limitations remain under heavy compression, weak audio quality, and cross-dataset generalization. Future improvements will focus on stronger domain adaptation, enhanced audio-forensics modules, and broader evaluation across diverse real-world datasets to further improve reliability and scalability.

In addition, this project demonstrates that a balanced design can be more practical than pursuing only maximum benchmark accuracy. By combining acceptable detection performance with low operational latency and deployable system components, True Sight Lens moves beyond a purely experimental model toward real-world usability. The system architecture is modular, allowing individual components such as the visual classifier, lip-sync analyzer, or fusion strategy to be upgraded without redesigning the entire pipeline. This extensibility is important for deepfake detection, where manipulation techniques evolve rapidly and static detectors degrade over time.

Overall, True Sight Lens provides a strong foundation for scalable media authenticity analysis in social and web environments. The present implementation validates the feasibility of multi-modal forensic screening in a user-facing application while clearly identifying technical gaps that guide future research. With continued improvements in cross-domain robustness, compression-aware training, and adaptive fusion, the framework can evolve into a more reliable and generalized defense layer against emerging deepfake threats.

ACKNOWLEDGEMENT

The authors acknowledge the support and facilities provided by REVA University, which contributed significantly to the successful completion of this project. The authors gratefully acknowledge Dr. P. Shyama Raju, Chancellor, of REVA University, for his visionary leadership and for fostering a conducive academic and research environment. The authors express their sincere appreciation to Dr. Ashwin Kumar U. M., Director, School of Computer Science and Engineering, for his valuable guidance and encouragement throughout the course of this project. The authors also extend their sincere thanks & deeply indebted to their project supervisor to Dr N P Nethravathi, Head of the Department, B.Tech in Computer Science & Engineering (Internet Of things & Cyber Security including Block Chain Technology), School of Computer Science and Engineering, for her continuous support, insightful feedback, constructive guidance and encouragement at every stage of the project. The authors also thank all the faculty members of the School of Computer Science and Engineering, REVA University, for their support and motivation throughout the program.

REFERENCES

[1] J. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.

[2] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4401–4410.

[3] Rössler et al., “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 1–11.

[4] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Long Beach, CA, USA, Jun. 2019, pp. 46–52.

[5] Dolhansky et al., “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, Jun. 2020.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 3207–3216.

[7] Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Hong Kong, China, Dec. 2018, pp. 1–7.

[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," in Proc. IEEE Int. Conf. Biometrics (ICB), Crete, Greece, Jun. 2019, pp. 1–8.

[9] Sabir et al., "Recurrent convolutional strategies for face manipulation detection in videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Long Beach, CA, USA, Jun. 2019, pp. 80–87.

[10] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 910–932, Aug. 2020.

[11] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and A. Morales, "Deepfakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131–148, Dec. 2020.