iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84339

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

BlinkNet: GRU-Based Temporal Analysis for Deepfake Video Detection

Gyara Dhanush[1], Dr. I. Lakshmi Manikyamba[2]
[1]*Post Graduate Student, M. Tech, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India*
[2]*Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India*

Abstract: ***The increasing realism of synthetic facial videos has reduced the reliability of detectors that depend only on visible artifacts in isolated frames. This paper presents BlinkNet, an explainable deepfake-detection framework that examines the spatial appearance and temporal kinematics of eye blinks. The system detects a face, localizes 68 facial landmarks, extracts normalized ocular crops, and computes the Eye Aspect Ratio (EAR) for each frame. Overlapping sequences of 20 frames are processed by a dual-stream Temporal-Spatial Physiological Blink Anomaly Network (TPBAN). A lightweight MobileNetV2 encoder models local visual inconsistencies, while a bidirectional gated recurrent unit models the forward and backward dynamics of eyelid motion. Temporal attention assigns a relevance weight to every frame and supports frame-level anomaly visualization. Training uses a multi-task objective for authenticity classification and blink-phase recognition, together with a class-weighted binary cross-entropy term to address the imbalance between genuine and manipulated sequences. On the FaceForensics++ c23 test partition, BlinkNet obtained 83.16% accuracy, 87.31% ROC-AUC, 96.02% average precision, and a 21.08% equal error rate. The manipulated class achieved 0.90 precision and 0.88 recall. The implementation processed video at approximately 52 frames per second on a consumer laptop GPU and was integrated into a Flask-based forensic dashboard. These results indicate that ocular dynamics can complement spatial evidence while improving efficiency and interpretability.***

Keywords*: Deepfake detection, eye-blink dynamics, Eye Aspect Ratio, BiGRU, MobileNetV2, temporal attention, digital forensics, explainable artificial intelligence.*

I. INTRODUCTION

Today, the art of generating fake videos has advanced beyond crude face-swaps to the generation of realistic video. Such a technology raises practical concerns regarding financial, security and evidential applications including news authenticity, forensic validation and digital identity security. Many popular deepfake detection methods examine individual video frames independently and rely on spotting artifact cues such as the visible presence of boundary blending between original footage and synthetic faces, unnatural texture, or an incongruent illumination. These cues still exist but can lose efficacy when a generator creates visually clean frames or the high visual frequency of these cues is lost due to video compression.

Humans also have additional ways of generating digital forensic evidence: For example, eye blinks are not single static events; rather, they are short physiological sequences that feature an orderly progression from open eyes to closed eyes and then back to the open position. Regardless of how visually authentic individual synthesized images may be, patterns in a person's eyelid's blink velocity, duration, synchronicity and regularity may be inconsistent between different times of observation. In prior work it has been established that eye blink frequencies can provide robust indications of the authenticity of multimedia content, and recurrent neural modelling approaches have shown that consistent temporal patterning can be observed in genuine footage. Our work (BlinkNet) addresses this observation more broadly by modelling both the visual characteristics of individual eyes in tandem with motion derived from the estimated eye region's geometry.

Our proposed approach adheres to four design principles: first, by using the image of an individual's eye as its primary input it filters out background noise thereby restricting the network's focus to the relevant parts of each image.

Second, we employ a shallow convolutional network to learn image features which can capture spatial artifacts as well as a bidirectional gated recurrent unit to model long-term temporal information. Third, we incorporated temporal attention mechanisms in our model to reveal the specific frames contributing most to a predicted probability. Fourth, to account for potential imbalance in dataset classification our model applies weighted training loss, and we demonstrate its ability to run efficiently in real-time on typical commodity computer hardware. The developed method not only yields a predictive probability score but also offers a visual trace indicating frames that influence the classification outcome, aiding human verification.

II. RELATED WORK

1) Spatial deepfake detection FaceForensics++ benchmark provides real clean videos and 4 families of manipulated videos: Deepfakes, Face2Face, FaceSwap and Neural Textures [2]. The common spatial baselines exploit high-capacity convolution networks (e.g., Xception) on this dataset to detect artifacts at the pixel level [9]. While high performance is achievable when there are artifacts like blending issues or frequency artifacts, the frame-by-frame independent analysis doesn't directly verify that facial movements make physiological sense.
2) Temporal and physiological features the researchers Li, Chang, and Lyu noticed that anomalous eye blinks can be indicative of fake videos [1]. Gera and Delp proposed combining convolutional features and recurrent networks for detecting temporal artifacts. Furthermore, facial landmark-based analysis of eye blinks has been further strengthened by the Eye Aspect Ratio, which was developed for real time eye state estimation [4]. Although the above studies motivate temporal physiology-based approaches, it's still required to be effective with video compression, require less computation, and generate explainable findings in the real-world usage of the detector
3) Lightweight and Explainable Modelling Inverted residual blocks and linear bottleneck units of MobileNetV2 make the model computationally efficient [5], therefore it is a good fit for feature encoding at the localized image level. With relatively few parameters compared to other standard LSTM models, the Gated Recurrent Units are chosen as the sequential model component [6]. To that extent, blinkNet integrates these and add temporal attention so that attention to each frame can be explicitly weighted for contributing to the final output. Different from any black-box video classifier model, blinkNet's prediction can be accompanied by visual timeline of attention weights.

III. PROPOSED METHODOLOGY

A. *System Overview*

The processing pipeline begins with an uploaded or streamed video. Frames are decoded using OpenCV, and a HOG-based face detector identifies the principal face. A 68-point landmark model locates the left eye at points 36–41 and the right eye at points 42–47. The ocular region is then represented in two complementary forms: normalized RGB eye crops and a scalar EAR signal. Overlapping 20-frame windows are supplied to TPBAN, which produces an authenticity probability, blink-phase estimates, and frame-level attention weights.

INSERT FIGURE 1 HERE

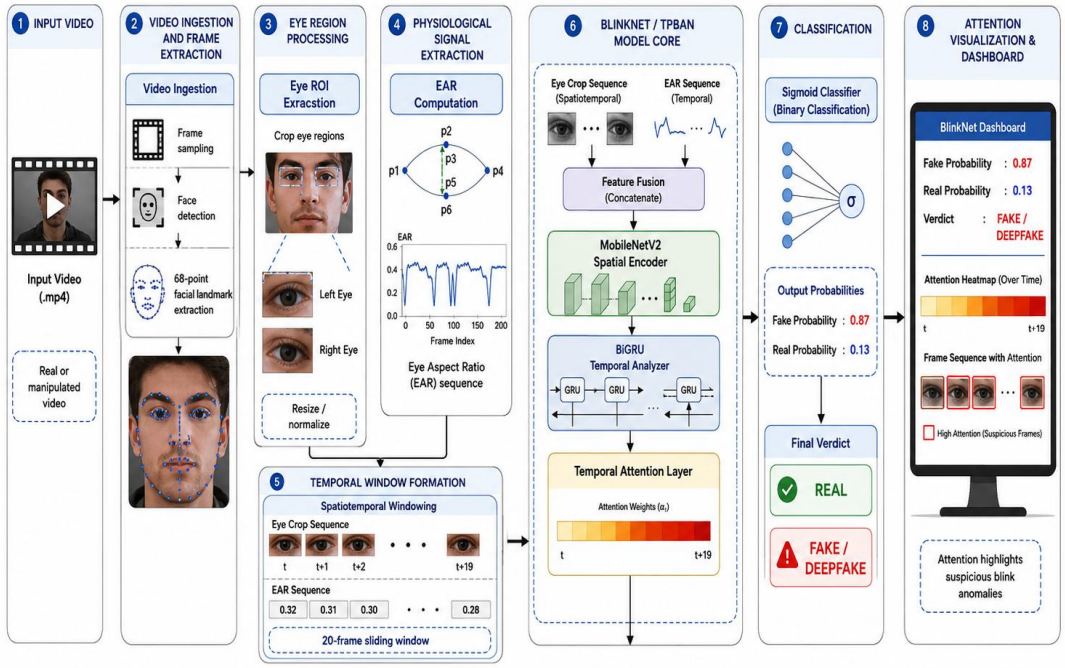


Fig. 1. End-to-end BlinkNet pipeline from video ingestion to classification and attention visualization.

B. Eye Localization and Physiological Signal Extraction

For each detected eye, six ordered landmark points describe the eyelid contour. EAR compares the vertical eyelid distances with the horizontal eye width and therefore provides a scale-normalized estimate of eye openness [4].

$$EAR = (\|p_2 - p_6\|_2 + \|p_3 - p_5\|_2) / (2\|p_1 - p_4\|_2) \quad (1)$$

The left- and right-eye ratios are averaged to form one EAR value per frame. A typical blink appears as a short decrease followed by recovery. Because the measure is based on relative geometry, it is less sensitive than raw pixel values to moderate changes in camera distance. When a face is temporarily missed because of motion blur or head movement, the implementation can reuse the most recent valid landmarks to preserve tensor dimensions; however, extended tracking failure causes the affected sequence to be rejected rather than treated as reliable evidence.

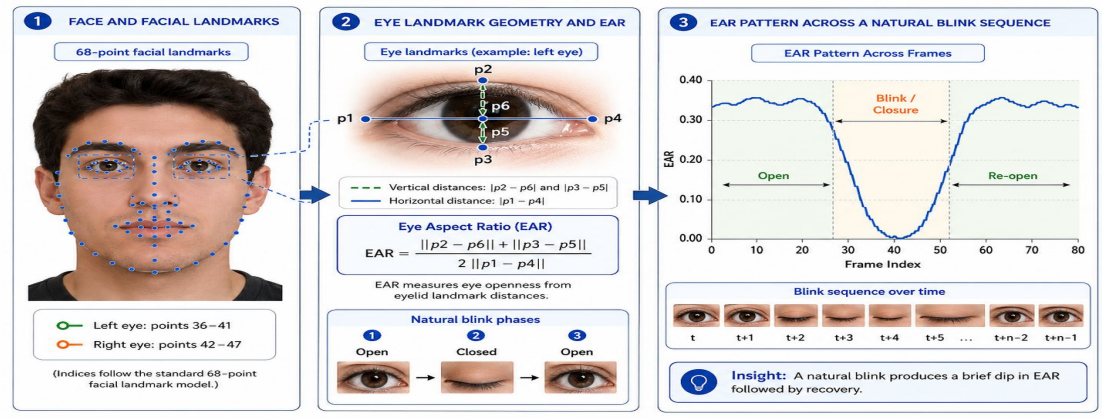


Fig. 2. Facial eye landmarks and the EAR pattern across a natural blink sequence.

C. Spatial-Temporal Feature Fusion

Each ocular crop is padded to retain the eyelid and eyelashes during full closure, converted from BGR to RGB, normalized to [0,1], and resized to 64 × 128 pixels. The frame encoder maps crop I_t to a compact spatial descriptor $s_t = f(MNet)(I_t)$. The descriptor is concatenated with the corresponding EAR value e_t, producing $x_t = [s_t; e_t]$. A bidirectional GRU processes the sequence in both chronological directions so that the representation of a frame includes the context before and after it.

$$h_t = [GRU\rightarrow(x_t); GRU\leftarrow(x_t)] \quad (2)$$

An attention layer converts each hidden state into a normalized importance coefficient. The final sequence representation is the weighted sum of hidden states, and a sigmoid classifier maps this representation to a deepfake probability.

$$a_t = \text{softmax}(v^T \tanh(Wh_t + b)), \quad z = \Sigma_t a_t h_t, \quad p = \sigma(w^T z + c) \quad (3)$$

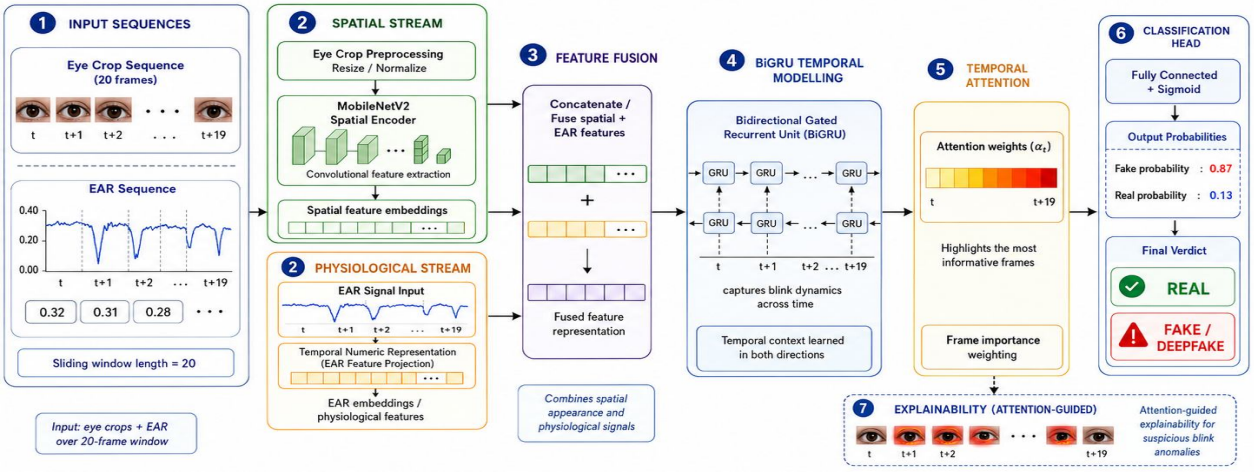


Fig. 3. Dual-stream TPBAN architecture with spatial encoding, EAR fusion, BiGRU modelling, and temporal attention.

D. Imbalance-Aware Multi-Task Training

The dataset windows are strongly imbalanced in favour of manipulated samples. A model trained with an unweighted objective could obtain deceptively high accuracy by over-predicting the majority class. BlinkNet therefore assigns a multiplier of 3.62 to errors on the minority genuine class. The authenticity loss is combined with an auxiliary five-class blink-phase loss covering pre-blink, closing, closed, opening, and post-blink states. The auxiliary task encourages the recurrent layers to learn the structure of a blink rather than only memorize manipulation-specific texture.

$$L = \alpha L\_authenticity + \beta L\_phase, \quad \alpha = 1.0, \beta = 0.3 \quad (4)$$

For authenticity prediction, the weighted binary cross-entropy increases the contribution of genuine samples to the gradient. This adjustment does not change the decision threshold; it changes the cost of errors during optimization and reduces majority-class collapse.

$$L\_authenticity = -(1/N)\Sigma_i[\omega_r(1-y_i)\log(1-p_i) + y_i\log(p_i)], \quad \omega_r = 3.62 \quad (5)$$

E. Inference and Visual Explanation

At inference time, a circular buffer stores 20 consecutive eye crops and EAR values. The buffer advances with a stride of 10 frames, allowing neighbouring windows to overlap. A probability above 0.5 is interpreted as manipulated media. Attention coefficients are aligned with the original frames and rendered as a temporal heatmap. High-weight frames are not treated as independent proof of manipulation; instead, they indicate where the model found the strongest inconsistency and help an analyst review the relevant eyelid transition.

IV. EXPERIMENTAL RESULTS

A. Dataset and Sequence Construction

Experiments used the FaceForensics++ benchmark [2]. The source collection contains 1,000 pristine videos, and the manipulation pipelines create 4,000 synthetic videos. The c23 compression level was selected because it reflects moderately compressed internet video and presents a realistic challenge for subtle physiological analysis. Videos were divided into training, validation, and test partitions before the extraction of overlapping 20-frame windows with a stride of 10.

Split	Windows	Purpose
Training	48,047	Parameter learning
Validation	9,532	Model selection
Testing	8,887	Final evaluation

TABLE I SEQUENCE WINDOW DISTRIBUTION

B. Implementation and Training Configuration

Training and evaluation were conducted on a laptop equipped with an AMD Ryzen 7 7735HS processor, 16 GB DDR5 memory, and an NVIDIA RTX 4050 GPU with 6 GB VRAM. The software environment used Python 3.11, PyTorch 2.x with CUDA 12.1, OpenCV, dlib, NumPy, scikit-learn, and Flask. AdamW optimization [8] was configured with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. A cosine OneCycle learning-rate schedule was applied over a maximum of 50 epochs with a 10% warm-up period. Automatic mixed precision reduced memory usage, gradients were clipped to an L_2 norm of 1.0, and an exponential moving average with decay 0.999 stabilized model selection.

C. Evaluation Metrics

Accuracy reports the proportion of all correct decisions, while precision and recall describe the reliability and coverage of each class. The F1-score summarizes the balance between precision and recall. ROC-AUC evaluates ranking performance across thresholds, average precision summarizes the precision-recall curve, and Equal Error Rate identifies the operating point at which false acceptance and false rejection rates are equal. EER is particularly useful for threshold-independent comparison in biometric and forensic systems.

V. RESULTS AND DISCUSSION

A. Overall Classification Performance

Table II summarizes the test-set performance. The ROC-AUC of 87.31% shows that the network generally assigns higher manipulation scores to synthetic sequences than to genuine sequences. The average precision of 96.02% is strengthened by the large manipulated class, while the EER of 21.08% reveals that threshold calibration remains important when false acceptance and false rejection carry similar cost.

Metric	Value
Accuracy	83.16%
ROC-AUC	87.31%
Average Precision	96.02%
Equal Error Rate	21.08%
Approx. inference speed	52 FPS

TABLE II GLOBAL TEST PERFORMANCE

The manipulated class achieved 0.90 precision and 0.88 recall. Thus, nine out of ten windows flagged as fake were correct, and the detector captured most manipulated windows in the test set. Performance on genuine media was weaker, with 0.60 precision and 0.66 recall. This gap is consistent with the difficulty of modelling natural variation in blinking, including differences caused by pose, eyewear, occlusion, fatigue, lighting, and compression.

Class	Prec.	Recall	F1	Support
Real	0.60	0.66	0.63	1,921
Fake	0.90	0.88	0.89	6,966
Weighted avg.	0.84	0.83	0.83	8,887

TABLE III CLASS-WISE CLASSIFICATION REPORT

Ground truth	Pred. real	Pred. fake
Real	1,276	645
Fake	852	6,114

TABLE IV CONFUSION MATRIX AT A 0.5 THRESHOLD

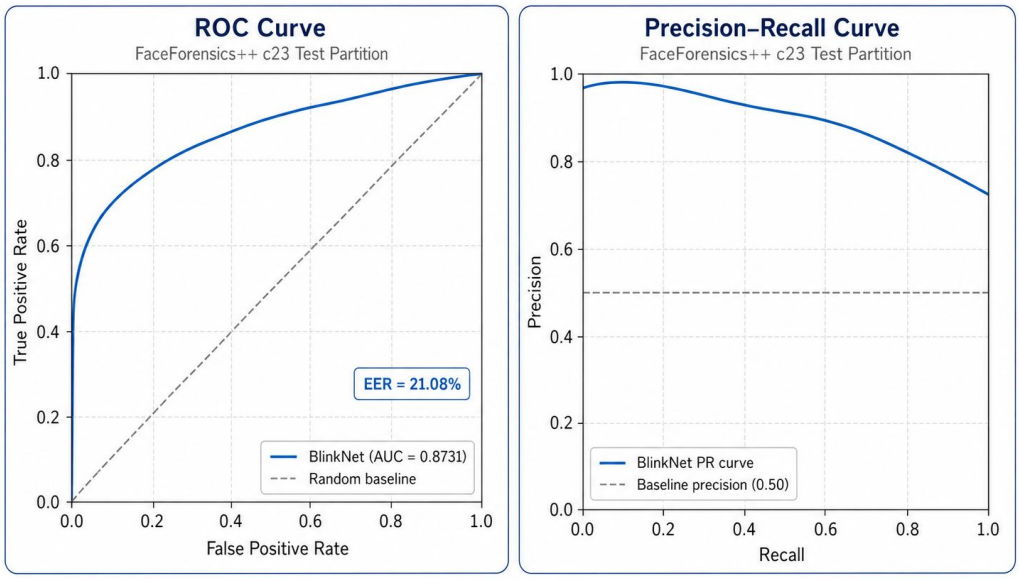


Fig. 4. ROC and precision-recall curves for the FaceForensics++ c23 test partition.

B. Effect of Class Weighting and Compression

The 4:1 imbalance between manipulated and genuine windows could allow a trivial majority-class strategy to appear competitive. The 3.62 genuine-class weight reduced this tendency and allowed the detector to identify 1,276 genuine windows correctly while retaining 88% recall on manipulated media. Nevertheless, 645 genuine windows were marked as fake, indicating that the minority class remains the principal area for improvement.

Compression also affects the two streams differently. H.264 c23 compression removes high-frequency texture around eyelashes, eyelid boundaries, and the iris. This can weaken the spatial encoder. EAR, however, is derived from landmark geometry and continues to provide a temporal signal when fine texture is degraded. The resulting ROC-AUC suggests that the combination is more robust than either cue would be alone, although an ablation study is required to quantify the independent contribution of each stream.

C. Explainability and Deployment Behaviour

The temporal attention layer converts the sequence decision into a frame-indexed explanation. In a natural blink, attention should be distributed around the closing and reopening transition. A sharp concentration on an isolated frame may reveal a synthetic discontinuity, landmark jump, or blending failure. The Flask dashboard displays the predicted label, confidence, EAR trace, and attention heatmap so that a reviewer can inspect the frames that contributed most strongly to the verdict.

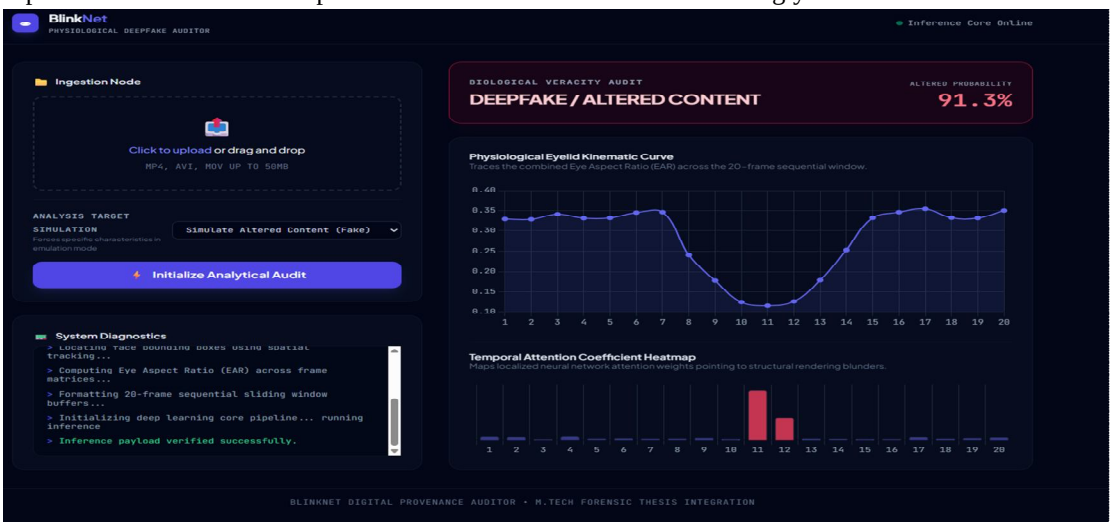


Fig. 5. Example deepfake result with frame-level temporal attention and anomaly localization.

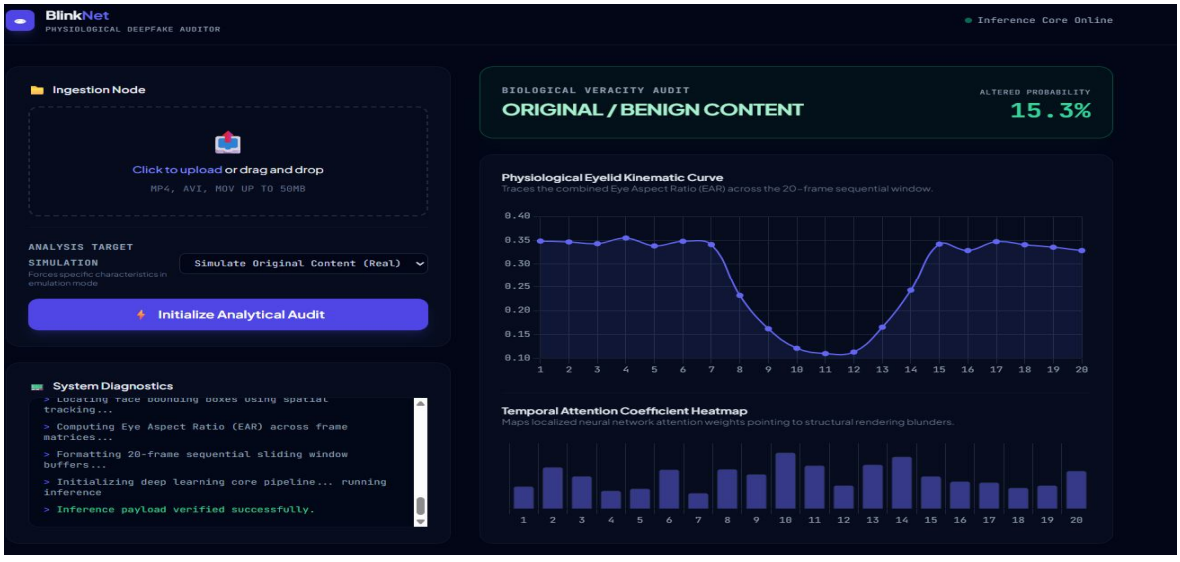


Fig. 6. Example genuine-video validation result in the BlinkNet web application.

The measured throughput of approximately 52 FPS on an RTX 4050 laptop GPU demonstrates the benefit of localized eye crops and a MobileNetV2 backbone. The speed value refers to the tested hardware and implementation; it should not be interpreted as a guaranteed rate on all devices, resolutions, or video codecs.

VI. LIMITATIONS AND FUTURE WORK

The current evaluation is limited to FaceForensics++ under the c23 protocol. A model may learn dataset-specific properties, and therefore cross-dataset tests on unseen manipulation methods are necessary before operational use. The pipeline also depends on reliable facial landmarks. Large head rotations, closed eyes for extended periods, masks, dark glasses, heavy motion blur, or very low resolution can reduce the quality of both the crop and EAR signal. A fixed 20-frame window may not cover every blink duration across different frame rates.

Future work should evaluate raw, c23, and c40 compression levels; perform explicit ablation studies for the spatial branch, EAR branch, BiGRU, attention, and auxiliary blink-phase objective; and calibrate decision probabilities across datasets. Landmark confidence can be incorporated into the attention mechanism so that tracking errors are not mistaken for synthetic anomalies. Additional physiological evidence, such as lip-audio synchronization [7], remote photoplethysmography, gaze continuity, and micro-expression dynamics, can be fused with blinking. Lightweight temporal convolution or transformer variants may further improve long-range modelling while retaining edge-device efficiency.

VII. CONCLUSION

BlinkNet demonstrates that eye-blink dynamics provide useful forensic information beyond static facial appearance. The proposed TPBAN architecture combines normalized ocular imagery, EAR-based physiology, MobileNetV2 feature extraction, bidirectional GRU sequence modelling, and temporal attention. On the FaceForensics++ c23 test partition, the system achieved 83.16% accuracy, 87.31% ROC-AUC, 96.02% average precision, and 21.08% EER, with 0.90 precision and 0.88 recall for manipulated sequences. Weighted training reduced majority-class bias, while the attention mechanism exposed the frames most responsible for a decision. Together with real-time performance on consumer hardware and a Flask interface, the framework provides a practical foundation for efficient and explainable deepfake screening. Broader cross-dataset validation and multi-physiological fusion remain essential before high-stakes deployment.

VIII. ACKNOWLEDGMENT

The authors acknowledge the support of the Department of Computer Science and Engineering, University College of Engineering, Science & Technology, Jawaharlal Nehru Technological University Hyderabad. The authors also thank the researchers who released the FaceForensics++ benchmark and the open-source tools used in this study.

REFERENCES

[1] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630763.

[2] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1–11, doi: 10.1109/ICCV.2019.00009.

[3] D. Güera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163.

[4] T. Soukupová and J. Čech, “Real-Time Eye Blink Detection Using Facial Landmarks,” in Proc. Computer Vision Winter Workshop, 2016, pp. 1–8.

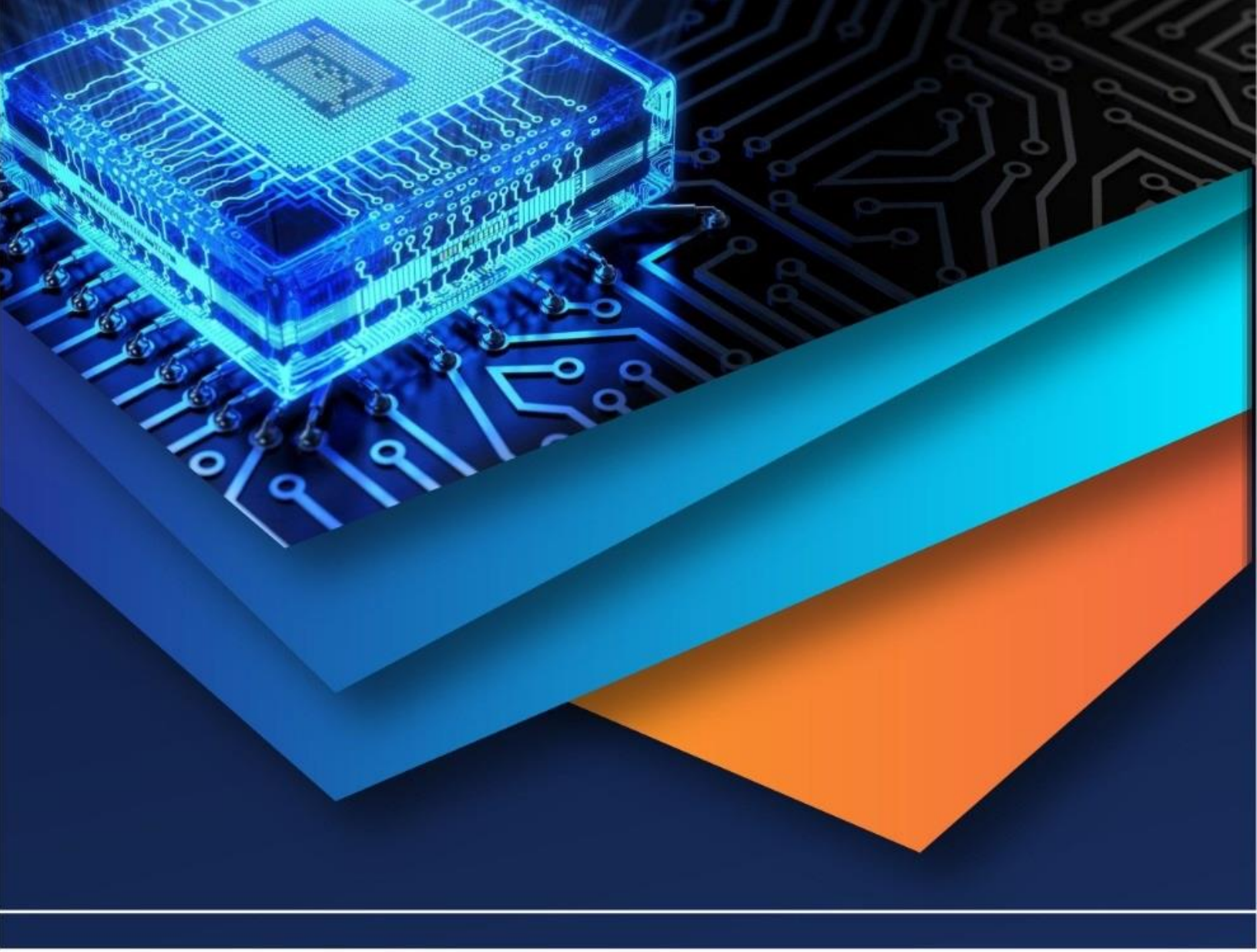
[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[6] K. Cho et al., “Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[7] J. S. Chung, A. Jamaludin, and A. Zisserman, “You Said That?: Synthesising Talking Faces from Audio,” in Proc. British Machine Vision Conf. (BMVC), 2017, doi: 10.5244/C.31.54.

[8] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[9] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258, doi: 10.1109/CVPR.2017.195.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY