Proceedings of the 6th International Conference on Inventive Computation and Information Technologies (ICICIT-2026)
DVD Part Number: CFP26UU7-DVD; ISBN: 979-8-3315-5010-3

A Bio-Kinematic Consistency Index Framework for Deepfake Video Detection

Sharon[1], Dr. N. Shyamala Devi[2]
[1]Research Scholar, Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies, Chennai, India
sharon@patriciancollege.ac.in
[2]Assistant Professor, Department of Computer Application, Vels Institute of Science, Technology & Advanced Studies, Chennai, India
shyamadevi@gmail.com

Abstract—Modern deepfakes increasingly suppress visible synthesis artifacts, while social-media processing further changes resolution, illumination, frame rate, and compression traces. This paper presents the Bio-Kinematic Consistency Index (BKCI), a lightweight and interpretable video-level framework that measures whether physiological color variation, regional facial micro-motion, mouth-face coupling, and their temporal relationships remain mutually consistent. The method converts regional evidence into physiological disorder, kinematic disorder, and perturbation-based temporal-instability descriptors and classifies the resulting vector using lightweight learners. On the held-out Celeb-DF test partition, BKCI-SVM obtained 77.46% accuracy, 78.45% F1-score, and 82.83% AUC. A fusion model increased AUC to 84.68%, 3.65 percentage points above the artifact-RF baseline. Under low-resolution recompression, 15-fps reduction, and low-light degradation, the selected BKCI variants retained AUC values of 80.45%, 81.70%, and 76.84%, respectively. These results indicate that bio-kinematic consistency is most useful as complementary forensic evidence; low illumination remains the principal failure condition.

Keywords **- Deepfake detection, video forensics, rPPG, optical flow, temporal perturbation, bio-kinematic consistency.**

I. INTRODUCTION

Deepfake generation through face swapping, reenactment, neural rendering, and diffusion-based synthesis threatens digital identity, biometric authentication, public communication, and multimedia forensics. Modern synthesis systems can alter facial appearance, expression, speech behavior, and identity while maintaining convincing visual realism. These manipulations are especially concerning where social-media platforms, financial services, authentication systems, or investigators rely on visual evidence.

Early deepfake detection approaches primarily focused on spatial artifacts such as blending boundaries, texture inconsistencies, abnormal color distributions, generator fingerprints, and compression traces. Although these methods perform well under controlled conditions, their reliability decreases as synthesis quality improves. Resizing, recompression, illumination changes, frame-rate reduction, and platform processing can further weaken their cues. Consequently, spatial artifact-based approaches often struggle to generalize across datasets and real-world deployment conditions.

Recent research has therefore shifted toward temporal, physiological, and semantic consistency analysis. Biological-signal approaches show that synthetic videos may not preserve coherent physiological patterns across facial regions. Temporal approaches instead identify inconsistencies in facial motion and synchronization. Physiological signals are weak and sensitive to illumination, pose, compression, and motion artifacts. Deep temporal models, meanwhile, often require large datasets and substantial computation to generalize.

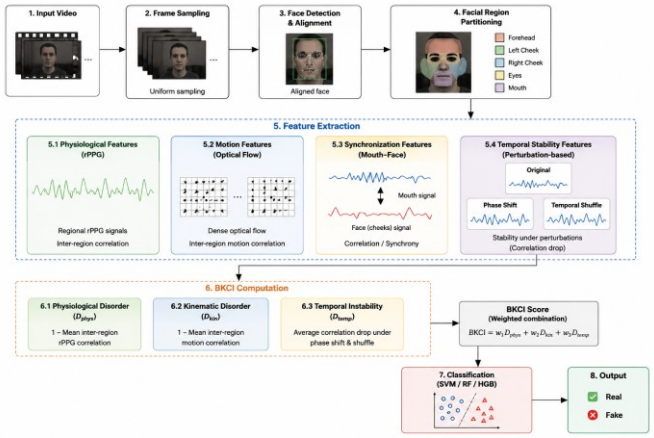

Fig. 1. Overview of the proposed Bio-Kinematic Consistency framework

These limitations highlight the need for lightweight and interpretable approaches that utilize evidence beyond visible artifacts while remaining robust under practical video-processing conditions. Authentic videos generally preserve coupled physiological behavior, coordinated facial micro-motion, and temporally stable regional relationships due to underlying biological and facial dynamics. Manipulated videos may preserve local realism while disrupting these subtle relationships across facial regions and over time.

To address these limitations, this paper proposes BKCI as a video-level consistency representation rather than a direct heart-rate estimator. BKCI measures agreement among normalized regional color traces, facial micro-motion, mouth-face coupling, and their response to controlled temporal perturbations. The scope is binary real-versus-fake detection for videos containing one sufficiently visible face; audio-only manipulation, fully occluded faces, extreme profiles, and attribution of the generating model are outside the present scope.

The main contributions of this work are summarized as follows:

- A Bio-Kinematic Consistency Index (BKCI) formulation that organizes physiological coherence, facial micro-motion synchrony, and mouth-face coupling into explicit disorder representations.
- A perturbation-based temporal stability mechanism that evaluates whether regional consistency patterns

remain meaningful under phase-shift and shuffle operations.

- A video-level evaluation against an artifact baseline and three lightweight BKCI classifiers, together with controlled tests for low-resolution recompression, frame-rate reduction, and low-light degradation.
- Experimental evidence demonstrating that bio-kinematic consistency provides lightweight, interpretable, and complementary forensic evidence beyond conventional artifact-based approaches.

II. LITERATURE REVIEW

Biological-signal-based deepfake detection gained significant attention after the work of FakeCatcher [1], which demonstrated that synthetic portrait videos often fail to preserve coherent physiological patterns across facial regions. The study utilized biological signal maps extracted from facial regions to identify manipulated videos and demonstrated that physiological inconsistencies can provide useful forensic evidence beyond conventional spatial artifacts. Similarly, DeepFakesON-Phys [2] investigated heart-rate-related features for deepfake detection and demonstrated that physiological measurements can contribute to forgery detection. Although these approaches introduced physiological evidence into deepfake forensics, their performance remains sensitive to illumination variation, motion artifacts, compression, and face-quality degradation. Furthermore, isolated physiological estimation may become unreliable for short or low-quality videos.

FaceForensics++ [3] established one of the most influential benchmarks for manipulated facial content and demonstrated that convolutional architectures can achieve strong performance under controlled compression settings. The study introduced multiple manipulation categories and enabled systematic evaluation of deepfake detectors under standardized conditions. Subsequently, Xception [5], based on depthwise separable convolutions, became one of the most widely adopted baselines because of its ability to learn local texture artifacts efficiently. EfficientNet [6] further improved the tradeoff between computational complexity and performance through compound scaling strategies. While these methods achieved strong in-domain performance, they frequently relied on dataset-specific visual artifacts that may not generalize effectively under recompression, resizing, or cross-domain evaluation.

Frequency-domain approaches investigated spectral artifacts introduced during synthesis and post-processing operations. Frequency-aware forgery detection [12] demonstrated that generative models may introduce abnormal frequency characteristics that are difficult to observe directly in the spatial domain. These approaches are attractive because generator artifacts may remain detectable even when spatial artifacts become less visible. However, frequency cues are often sensitive to compression settings, platform-specific processing, and resolution changes, thereby limiting robustness across datasets and deployment environments.

Temporal modeling approaches attempted to capture inconsistencies across consecutive frames rather than analyzing frames independently. CNN-LSTM architectures, recurrent models, optical-flow approaches, and video transformers were introduced to learn temporal dependencies and motion relationships across manipulated videos. Temporal consistency is particularly important because authentic facial behavior involves coordinated motion among facial components such as the mouth, cheeks, eyes, and head regions. However, these approaches often require substantial computational resources and large-scale training data to generalize effectively.

Recent transformer-based approaches further explored long-range temporal and spatial dependencies. Vision Transformers (ViT) [7] demonstrated strong representation learning capabilities through attention mechanisms, while TimeSformer [8] introduced space-time attention for video understanding tasks. These architectures are attractive for deepfake detection because manipulations may introduce subtle inconsistencies distributed across both spatial and temporal dimensions. Nevertheless, transformer architectures remain computationally demanding and frequently require extensive training data to achieve stable performance.

Benchmark and representation studies broadened the field beyond single-cue detectors. Celeb-DF [4], DFDC [14], and DeeperForensics-1.0 [15] introduced increasingly challenging identities, synthesis pipelines, and real-world perturbations. Head-pose inconsistency [9], MesoNet [10], Face X-ray [11], and frequency-aware learning [12] target geometric, compact spatial, blending-boundary, and spectral evidence, respectively. The survey in [13] organizes these manipulation and detection families, while rPPG principles [16] and optical-flow estimation [17] provide the signal-processing foundations used by BKCI.

Recent work has increasingly emphasized generalization rather than in-domain accuracy. LipForensics [18] models high-level mouth dynamics, RealForensics [19] learns temporally coherent representations from real videos, self-blended images [20] reduce dependence on a particular manipulation generator, and UCF [21] separates common forgery cues from method-specific features. GM-DF [22] extends this direction to unified multi-scenario training, while Deepfake-Eval-2024 [23] demonstrates the substantial performance loss that contemporary detectors can experience on multilingual, in-the-wild media. These studies reinforce two design requirements adopted here: temporal evidence should be evaluated at the video level, and a detector should not rely on a single brittle artifact family.

Unlike existing physiological or temporal deepfake detection methods, the proposed BKCI framework integrates physiological coherence, facial micro-motion synchrony, mouth-face coupling, and perturbation-based temporal stability into a unified and interpretable video-level representation. To the best of our knowledge, no prior framework combines these complementary consistency measures within a single lightweight forensic descriptor.

III. PROPOSED METHOD

Figure 1 summarizes the processing chain, and Fig. 2 shows the three disorder families retained in the final representation. Let a video

provide T valid sampled frames and R=5 facial regions: forehead, left cheek, right cheek, mouth, and eye. All descriptors are computed per video; the classifier therefore receives one vector per video rather than independent frame labels.

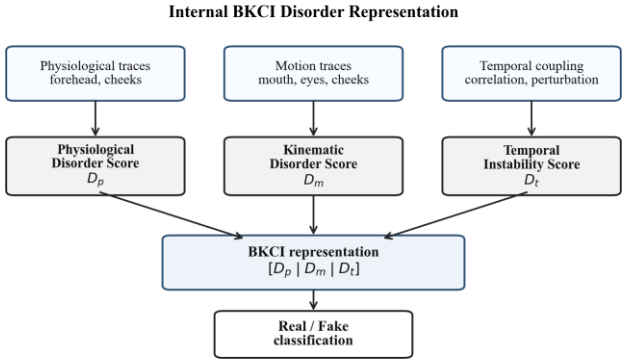

Fig. 2. Internal BKCI disorder representation showing physiological disorder, kinematic disorder, and temporal instability components.

A. Dataset, Face Processing, and Video-Level Protocol

The evaluation uses 1,180 balanced Celeb-DF videos: 590 authentic and 590 manipulated. As listed in Table I, 826 videos form the training partition, 177 the validation partition, and 177 the held-out test partition. Sampling and feature extraction are performed within each video, and performance is reported only after aggregating evidence to a single video-level vector. The validation partition is reserved for model selection; the test partition is used for the final metrics reported in Tables II and III.

Each video contributes at most 96 frames, sampled uniformly over its full duration. Frames wider than 320 pixels are resized to a width of 320 while preserving aspect ratio. The largest frontal face is detected with the OpenCV Haar cascade using scaleFactor=1.1, minNeighbors=5, and a minimum face size of 48 x 48 pixels. A sample is retained only when at least eight frames are available and faces are detected in at least eight frames. The detected face is partitioned into fixed relative forehead, left-cheek, right-cheek, mouth, and eye regions.

For region r and frame t, g_r(t) is the mean green-channel intensity. Regional summary features include the mean, standard deviation, extrema, skewness, kurtosis, and dominant frequency. Dominant frequency is estimated with Welch's method using at most 64 samples and the 0.6-4.0 Hz band. For motion extraction, each region is converted to grayscale and resized to 32 x 32 pixels. Dense Farneback flow uses pyr_scale=0.5, two pyramid levels, an 11-pixel window, two iterations, poly_n=5, and poly_sigma=1.1. These signals are forensic descriptors inspired by rPPG [16], not clinical pulse measurements.

B. Physiological and Kinematic Disorder

Pairwise physiological agreement is c_ij=|corr(z_i,z_j)| for the non-mouth facial regions. The physiological-disorder component is D_p=1-(1/|P|) sum_(i,j in P)c_ij, where P is the set of valid region pairs. D_p approaches zero when regional color variations are coherent and increases when their temporal behavior disagrees. Pairwise differences and distribution summaries are retained so the classifier can distinguish one anomalous region from a globally weak signal.

Dense optical flow between adjacent aligned frames provides a regional motion magnitude m_r(t) and direction summary. After robust temporal normalization, the pairwise motion agreement q_ij=|corr(m_i,m_j)| yields D_k=1-(1/|K|) sum_(i,j in K)q_ij. This term captures loss of coordinated facial micro-motion without requiring a large end-to-end video network.

Mouth-face coupling is measured by comparing mouth motion m_mouth(t) with the aggregate non-mouth facial motion m_face(t). The coupling disorder is D_m=1-|corr(m_mouth,m_face)|. It is informative during speech and expression changes, but it is treated as one component rather than definitive evidence because silent or nearly static clips naturally provide weak mouth motion.

C. Temporal Perturbation and Stability

Temporal validity is tested by disrupting alignment rather than assuming that every correlation is meaningful. Skin-region phase perturbation uses a circular shift of max(2,floor(fps/5)) frames; mouth-face motion uses max(1,floor(fps/6)). Temporal shuffling uses a fixed random seed of 17. For a signal pair, the perturbation response is the reduction in absolute correlation after the shift or shuffle. A large positive reduction indicates an alignment-dependent relationship, whereas a small reduction indicates that the apparent agreement survives temporal destruction and is weak forensic evidence.

The temporal-instability term is D_t=1-mean(Delta) after scaling valid responses to [0,1]. Phase shifting tests sensitivity to timing, whereas shuffling tests sensitivity to temporal order. The two perturbations address different failure modes and prevent a high but temporally uninformative correlation from dominating BKCI.

D. BKCI Vector and Classification

The core BKCI vector is b=[D_p,D_k,D_m,D_t] and is extended with facial-symmetry differences, regional disagreement statistics, temporal means and variances, and luminance-normalized summaries. The representation retains each component instead of collapsing all evidence into one opaque scalar. An analyst can therefore see whether color inconsistency, motion disagreement, mouth coupling, or temporal instability primarily drives a decision.

The evaluated BKCI classifier uses 141 retained base and robustness-oriented feature columns. Missing values are imputed with the training-set median, and StandardScaler parameters are fitted on the training data only. BKCI-RF uses 500 trees, balanced-subsample class weights, min_samples_leaf=2, n_jobs=1, and random_state=42. BKCI-SVM uses an RBF kernel with C=3.0, gamma='scale', balanced class weights, probability calibration enabled, and random_state=42. Both models are trained on 813 valid training videos; 173 valid validation videos support model comparison. Test predictions use a probability threshold of 0.5.

E. Robustness Protocol, Metrics, and Complexity

The robustness protocol transforms held-out videos without retraining. Low-resolution recompression downsizes each frame to at most 360 pixels in width with area interpolation and writes MP4V video. Frame-rate reduction retains frames at stride round(original fps/15) and writes the output at 15 fps. Low-light processing applies I'=0.60I-8 with OpenCV saturation. These deterministic transformations isolate spatial loss, temporal subsampling, and illumination loss, respectively.

Accuracy, F1-score, and AUC are reported because they answer different questions. Accuracy measures thresholded correctness on the balanced split, F1-score balances precision and recall for the manipulated class, and AUC measures ranking quality independently of one operating threshold. AUC is emphasized for robustness analysis because degradation may shift score calibration even when ranking remains informative.

Computationally, feature extraction scales linearly with the number of sampled frames, while classifier inference operates on a fixed-length video vector. The main practical cost is face alignment and dense optical flow; the classical classifier adds comparatively little

overhead. This design targets screening and forensic triage rather than generator attribution or clinical physiological measurement.

F. Scope, Limitations, and Practical Applications

Scope: BKCI addresses binary real-versus-fake classification for videos that contain one sufficiently visible and trackable face. It analyzes visual physiological and kinematic consistency at video level. It does not detect audio-only manipulation, identify the synthesis method, attribute the creator, or estimate a clinical heart rate.

Limitations: The present evaluation uses one Celeb-DF split and 173 valid held-out test videos after quality filtering. Consequently, the results do not establish universal cross-dataset generalization or statistical superiority across identities and manipulation families. Performance also depends on reliable face tracking, adequate facial resolution, and usable illumination. Extreme pose, occlusion, short static clips, and low light can weaken the regional color and motion signals. Dense optical flow remains the principal computational cost.

Practical applications: BKCI is intended as an interpretable secondary signal for forensic triage, content-moderation review, and integrity checks in identity-verification workflows. Its component scores can help an analyst determine whether a decision is driven by physiological disagreement, abnormal facial motion, weak mouth-face coupling, or temporal instability. Because these cues can also be degraded by benign capture conditions, BKCI should be combined with artifact, provenance, audio, or human-review evidence rather than used as stand-alone proof of manipulation.

IV. RESULTS AND DISCUSSION

The BKCI pipeline was implemented in Python and evaluated on the balanced video-level partitions in Table I. The analysis asks three questions: whether BKCI ranks manipulated videos reliably, whether it contributes evidence beyond an artifact baseline, and how strongly common degradations alter its behavior.

The full dataset contains 590 real and 590 fake videos. Feature quality checks retained 813 training, 173 validation, and 173 test videos. Four of the 177 test videos were excluded because fewer than eight usable face detections were available. The valid test set contains 85 authentic and 88 manipulated videos. This accounting prevents failed face tracking from being silently counted as physiological disorder.

Table I gives the exact split. The use of separate validation and test partitions is important for a lightweight model comparison: model selection is separated from the final evaluation, and all reported metrics are calculated at video level.

TABLE I. DATASET DISTRIBUTION USED FOR EXPERIMENTAL EVALUATION

Dataset Split	Real Videos	Fake Videos	Total
Training	413	413	826
Validation	88	89	177
Testing	89	88	177
Total	590	590	1180

Table II and Fig. 3 compare the reported models. BKCI-SVM improves AUC from 81.03% for Artifact RF to 82.83% (+1.80 percentage points), although its accuracy is 1.07 points lower and its F1-score is 0.20 points lower. This divergence means BKCI-SVM ranks samples better but does not use the best fixed operating threshold. The fusion model reaches the highest AUC, 84.68%, improving by 3.65 points over Artifact RF and 1.85 points over BKCI-SVM. The gain supports complementarity between artifact and bio-kinematic evidence, not wholesale replacement of artifact cues.

The fusion model's higher AUC is consistent with complementary error patterns. Artifact RF responds to spatial texture, blending, and compression traces, whereas BKCI responds to cross-region physiological and motion relationships. A video that suppresses visible artifacts may still disturb bio-kinematic consistency; conversely, heavy post-processing may weaken BKCI while leaving some artifact evidence. Combining the two feature families therefore improves ranking when either family alone is ambiguous. This explanation is supported by the AUC gain, although per-video error-overlap analysis would be required to quantify the complementarity directly.

BKCI-SVM also achieves a higher AUC than BKCI RF. A plausible reason is that the standardized BKCI descriptors form a compact continuous feature space in which a margin-based boundary preserves relative score ordering. Random forests partition that space into piecewise-constant regions and can be less stable when the sample size is modest and correlated regional summaries create many similar split candidates. This interpretation explains the observed ranking difference, but it should be confirmed through repeated runs and hyperparameter-controlled ablation rather than treated as a universal advantage of SVM.

TABLE II. PERFORMANCE COMPARISON OF EVALUATED DEEPFAKE DETECTION METHODS

Method	Accuracy (%)	F1 Score (%)	AUC (%)
Artifact RF	78.53	78.65	81.03
BKCI RF	76.30	78.31	81.56
BKCI-SVM	77.46	78.45	82.83
Fusion Model	76.30	78.31	84.68

BKCI RF and the fusion model share the same reported thresholded accuracy and F1-score (76.30% and 78.31%) but differ by 3.12 AUC points. The result again exposes a calibration issue: ranking improves without a corresponding gain at the selected threshold. In deployment, the operating threshold should therefore be calibrated on validation data for the required false-positive cost rather than copied unchanged across environments.

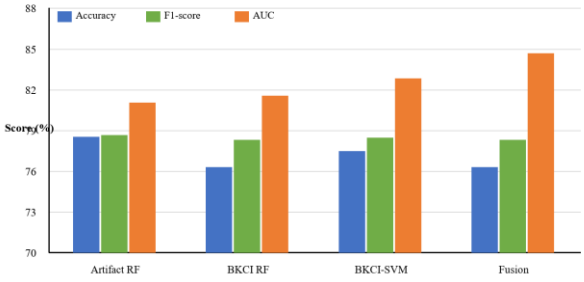

Fig. 3. Accuracy, F1-score, and AUC for the evaluated methods (values from Table II).

Table III and Fig. 4 report the strongest saved BKCI configuration for each condition. BKCI-SVM is used for the original, low-resolution, and low-light rows; BKCI-RF is used for the 15-fps row. Relative to the original BKCI-SVM test result, low-resolution recompression reduces accuracy by 2.89 points and AUC by 2.38 points. The 15-fps BKCI-RF result retains an AUC of 81.70%; because it uses a different classifier, it is interpreted as condition-level robustness evidence rather than a paired SVM degradation estimate.

TABLE III. BKCI ROBUSTNESS PERFORMANCE UNDER PRACTICAL VIDEO CONDITIONS

Condition	Accuracy (%)	AUC (%)	Model
Original Test Set	77.46	82.83	BKCI-SVM
Low Resolution + Recompression	74.57	80.45	BKCI-SVM
FPS Reduction	73.99	81.70	BKCI-RF

Condition	Accuracy (%)	AUC (%)	Model
Low Light	55.81	76.84	BKCI-SVM

Low light is the dominant failure condition: accuracy falls by 21.65 points, from 77.46% to 55.81%, while AUC falls by 5.99 points, from 82.83% to 76.84%. The smaller AUC drop relative to accuracy suggests that luminance loss shifts the score distribution in addition to weakening discrimination. This behavior is consistent with BKCI's dependence on subtle regional color variation and motivates explicit low-light quality control and threshold recalibration.

Low illumination affects BKCI at the signal source. Fewer detected photons increase sensor noise, reduce color-channel signal-to-noise ratio, and compress the small green-channel variations used as physiological descriptors. Exposure control, denoising, and video compression can further distort temporal color relationships, while motion blur and weak facial contrast reduce face alignment and optical-flow accuracy. These effects jointly explain why low light produces a larger accuracy loss than recompression or frame-rate reduction. The remaining AUC of 76.84% indicates that some ordering information survives, but the threshold requires recalibration.

The results support three bounded conclusions. First, BKCI is complementary to artifact cues because fusion gives the best AUC. Second, the selected BKCI variants retain AUC values above 80% under recompression and 15-fps processing, supporting further study on processed social-media video. Third, illumination is a material limitation. Practical uses include forensic triage, a secondary signal in content-moderation pipelines, and an interpretable corroborating score for human analysts; BKCI should not be treated as sole proof of manipulation.

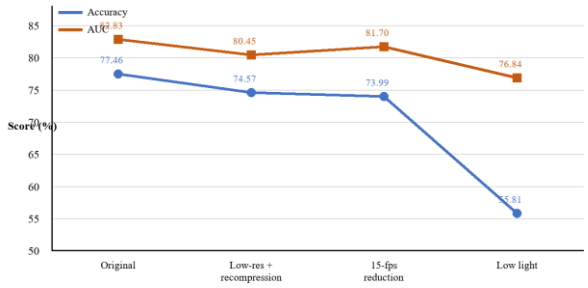

Fig. 4. Performance of the selected BKCI variants under practical video degradations (values from Table III).

Held-out threshold behavior is summarized by the confusion matrix in Table IV. At the 0.5 threshold, BKCI-SVM correctly classified 63 of 85 authentic videos and 71 of 88 manipulated videos. It produced 22 false positives and 17 false negatives, yielding 77.46% accuracy and 78.45% F1-score on 173 valid videos.

TABLE IV. BKCI-SVM HELD-OUT CONFUSION MATRIX

Actual class	Pred. real	Pred. fake
Authentic	63	22
Manipulated	17	71

Bootstrap validation quantifies sampling uncertainty. Table V reports 95% AUC confidence intervals from the saved per-video predictions. The original-test interval is 76.06-88.65%. The intervals remain above chance for all three degradation conditions, but their width reflects the modest number of valid videos.

TABLE V. BOOTSTRAP 95% AUC CONFIDENCE INTERVALS

Condition	n	AUC (%)	95% CI (%)
Original (SVM)	173	82.83	76.06-88.65
Low-res. (SVM)	173	80.45	73.41-86.71
15 fps (RF)	173	81.70	75.07-87.76
Low light (SVM)	172	76.84	69.18-83.82

Paired testing tempers the performance claim. Against Artifact RF on the 173 shared videos, BKCI-SVM improved AUC by 1.28 points, but the bootstrap interval for the difference was -5.03 to 7.52 points and the permutation p-value was 0.698; McNemar's p-value was 0.720. The corresponding comparison with the fusion model was also non-significant (AUC-difference p=0.505; McNemar p=0.868). Therefore, the experiments support complementarity and interpretability, not statistically proven superiority.

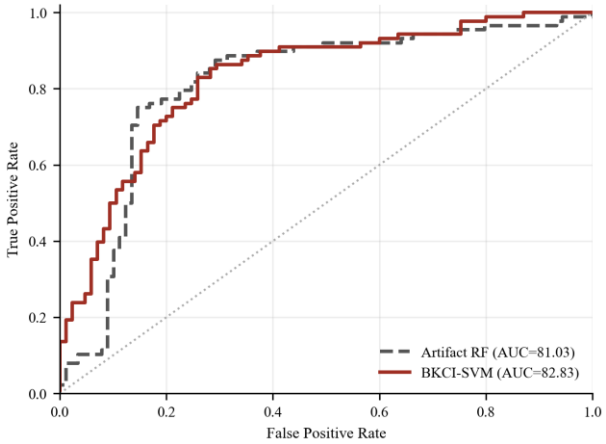

Fig. 5. Held-out ROC curves for Artifact RF and BKCI-SVM on the 173 shared valid videos.

Figure 6 provides a qualitative preprocessing check using a representative authentic Celeb-DF frame. The boxes correspond to the exact relative regions used by the feature extractor. This screenshot validates region placement; it is not presented as a detector-output heatmap.

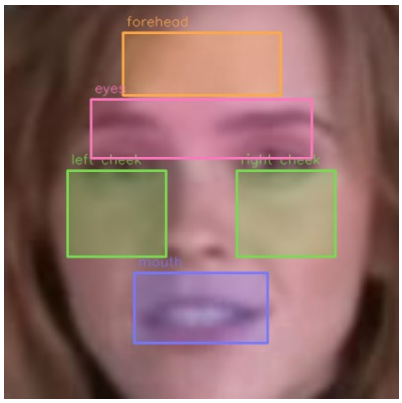

Fig. 6. Representative authentic frame showing the fixed facial regions used for BKCI extraction.

V. CONCLUSION AND FUTURE WORK

This paper presented BKCI, a lightweight video-level framework that converts regional color coherence, facial micro-motion, mouth-face coupling, and perturbation-based temporal stability into interpretable disorder descriptors. On 173 valid held-out Celeb-DF videos, BKCI-SVM achieved 77.46% accuracy, 78.45% F1-score, and 82.83% AUC; fusion with artifact evidence produced the best AUC of 84.68%. Bootstrap validation placed the BKCI-SVM AUC between 76.06% and 88.65%. The paired comparisons did not establish statistically significant superiority, so the defensible

outcome is complementarity and interpretability. Low-resolution BKCI-SVM retained 80.45% AUC, 15-fps BKCI-RF retained 81.70% AUC, and low-light BKCI-SVM retained 76.84% AUC, identifying illumination as the principal observed weakness. The present scope is limited to binary detection of videos with a sufficiently visible, trackable face; it does not address audio-only manipulation, generator attribution, extreme pose or occlusion, clinical pulse estimation, or universal cross-dataset generalization. Future work will evaluate repeated subject-disjoint and cross-dataset trials, improve low-light normalization and face-quality gating, test additional manipulation families, release complete code and configurations, and profile real-time throughput. In practice, BKCI is best used as an interpretable secondary signal for forensic triage or content-moderation review rather than as stand-alone proof of manipulation.

REFERENCES

[1] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 3965-3980, 2021.
[2] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-Phys: DeepFakes detection based on heart rate estimation," Proc. AAAI Workshop on Artificial Intelligence for Cyber Security, 2020.
[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," Proc. IEEE/CVF ICCV, pp. 1-11, 2019.
[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," Proc. IEEE/CVF CVPR, pp. 3207-3216, 2020.
[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. IEEE CVPR, pp. 1251-1258, 2017.
[6] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proc. ICML, pp. 6105-6114, 2019.
[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Proc. ICLR, 2021.
[8] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" Proc. ICML, pp. 813-824, 2021.
[9] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," Proc. IEEE ICASSP, pp. 8261-8265, 2019.
[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," Proc. IEEE WIFS, pp. 1-7, 2018.
[11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," Proc. IEEE/CVF CVPR, pp. 5001-5010, 2020.
[12] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," Proc. ECCV, pp. 86-103, 2020.
[13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131-148, 2020.
[14] B. Dolhansky et al., "The Deepfake Detection Challenge dataset," arXiv:2006.07397, 2020.
[15] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," Proc. IEEE/CVF CVPR, pp. 2889-2898, 2020.
[16] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," IEEE Trans. Biomed. Eng., vol. 64, no. 7, pp. 1479-1491, 2017.
[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proc. IJCAI, pp. 674-679, 1981.
[18] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "LipForensics: A temporal approach to detecting deepfakes," Proc. IEEE/CVF CVPR, 2021.
[19] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," Proc. IEEE/CVF CVPR, 2022.
[20] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," Proc. IEEE/CVF CVPR, 2022.
[21] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," Proc. IEEE/CVF ICCV, 2023.
[22] Y. Lai, Z. Yu, J. Yang, B. Li, X. Kang, and L. Shen, "GM-DF: Generalized multi-scenario deepfake detection," arXiv:2406.20078, 2024.
[23] N. A. Chandra et al., "Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," arXiv:2503.02857, 2025.