Frame-to-clip CNN Baseline for Video Deepfake Detection in English and Arabic

Rawad Abdulghafor [1,*], Yousuf Al Husaini [1], Mohammad Abrar [1], Abdullah Said Al-Aamri [2], Mohammed Abdulla Salim Al Husaini [1], and Wasin AlKishri [1]

[1] Faculty of Computer Studies (FCS), Arab Open University–Oman, Muscat, Oman
[2] College of Economics, Management and Information Systems, University of Nizwa, Nizwa, Oman
Email: rawad.a@aou.edu.om (R.A.); yousufnaser@aou.edu.om (Y.A.H.); abrar.m@aou.edu.om (M.A.); abdullah.alaamri@unizwa.edu.om (M.S.A.A.); mohammed.h@aou.edu.om (M.A.S.A.H.); wasan.k@aou.edu.om (W.A.)
*Corresponding author

***Abstract*—Deepfakes are becoming more authentic, which lowers confidence in digital media. This paper presents a lightweight and reproducible baseline for video deepfake detection that is tested on the English and Arabic subsets of the PolyGlotFake dataset. The method samples frames from each video and classifies them with a ResNet-18 model, and combines frame predictions to make a clip-level prediction. The proposed method can achieve an accuracy of 97.75% and a macro-Area Under the Curve (AUC) of 0.999, and a Matthews Correlation Coefficient (MCC) of 0.970 on a person-disjoint test split, demonstrating strong class separability and low computational cost. This work offers a grounded basis of practicality in bilingual video deepfake detection, through the provision of an open experimental pipeline and competitive functionality, as well as a basis for future temporal and multimodal extensions.**

***Keywords*—deepfake detection, video forensics, convolutional neural networks, English, Arabic**

I. Introduction

Cases of photorealistic video deepfakes are now more accessible, given that generative models have been developed, which has resulted in undermined trust and the ability to impersonate easily and on a larger scale [1]. Benchmark corpora have helped with standardized evaluation, and Rössler *et al*. [2] were the first to go ahead with facial manipulation detection and inter-method comparison. The DeepFake Detection Challenge (DFDC) data set further constituted additional training and benchmarking to increasingly realistic environments that highly value generalization [3]. Celeb-DF pushed the concept of visual realism and compression artifacts a step higher and exposed the vulnerability of the detectors trained on more basic distributions [4].

With these resources, cross-dataset generalization is still challenging, and WildDeepfake shows the systematic decrease in performance with actual "in-the-wild-media" [5]. Combination pipelines: Hybrid pipelines are pipelines that use powerful Convolutional Neural Network (CNN) backbones and Vision Transformers but are compute-intensive and dataset-conditional [6]. The orthogonal direction exploits the visual speech/lip dynamics to enhance the resilience to distribution shifts by modeling feasible facial movement [7]. The other complementary strand involves the use of physiological cues and DeepRhythm demonstrated that attentional heartbeat rhythms can reveal synthesis discrepancies of manipulated faces [8]. More recent surveys highlight that Out-of-Distribution (OoD) evaluation and robustness should be at the core of the design and reporting of methods instead of being an add-on [9]. Architecturally, there are simple frame-based CNN baselines, temporal (Recurrent Neural Networks/3D-CNNs/Transformers), and hybrids which trade off accuracy, compute, and stability across datasets [9]. Similar remote Photoplethysmography (rPPG)-based research goes an extra mile to apply biological cues to detect fake or even identify sources of a given synthesizer, emphasizing aspects that modern synthesis pipelines are unable to replicate consistently [10].

It is in this context that we build a computationally efficient video baseline on PolyGlotFake (English/Arabic) using frame sampling, ResNet-18 classification in MATLAB, and probability averaging to make clip-level decisions, following in the footsteps of frame-to-clip pipelines on large benchmarks. We also introduce fine-grained measures and pay attention to inter-breaks and inter-rater reliability too to minimize leakage of identity and measure strength; these are both encouraged by the new suggestions of the forensics literature. This countermeasure paper is a reaction to broader suggestions of more resilient countermeasures and verification infrastructure as deepfakes continue to propagate all over the globe.

The primary goal of this paper is to develop and evaluate a computationally efficient and reproducible video deepfake detection baseline suitable for bilingual scenarios involving English and Arabic content. Specifically, this study aims to:

Manuscript received October 17, 2025; revised December 10, 2025; accepted April 8, 2026; published November xx, 2026.

(1) Assess whether a lightweight frame-to-clip CNN pipeline based on ResNet-18 can achieve competitive performance without complex temporal models;
(2) Establish a person-disjoint evaluation protocol to reduce identity leakage and better reflect real-world deployment conditions; and
(3) Provide a transparent and replicable experimental framework that can serve as a reference baseline for future extensions, including temporal modeling and audio-visual fusion.

Through these objectives, the paper seeks to answer the research question: Can a simple, resource-efficient frame-based approach deliver robust and reliable video deepfake detection performance in bilingual settings.

II. Literature Review

The face forgery field has advanced significantly with the availability of large-scale public datasets. Benchmark datasets such as DeeperForensics-1.0 provide standardized evaluation frameworks that enable reproducible comparisons and support the development of robust detection methods across diverse manipulation and compression scenarios [11]. Furthermore, Celeb-DF introduces higher visual realism and more challenging artifacts, exposing the limitations of detectors trained on simpler datasets [4]. In addition, WildDeepfake extends evaluation to real-world, unconstrained media, highlighting the challenges of cross-dataset generalization and performance degradation in practical scenarios [5]. These come along with WildDeepfake choosing in-the-wild data scraped off the internet to show cross-dataset generalization errors, where the state-of-the-art systems are likely to plummet [5].

The other tool is DeeperForenshics-1.0, which added real-world post processing perturbation that builds on compression, noise, and lighting to work better and more efficient [11]. Another contribution to the body of knowledge made by ForgeryNet is that it generates encouraging results due to the unprecedented scale, which promotes the growth of holistic forensics rather than clip-level labels [12]. KoDF described distribution changes as associated with demographics, language, and capture pipelines, and the role multi-dataset training and evaluation play in the importance of multi-dataset training and evaluation [13]. In addition to these datasets, DeepfakeBench standardized protocols, data management and metrics with a wide range of methods and datasets to minimize apples-to-oranges comparisons [14].

Methodologically, early detectors treated videos as sets of independent frames processed by 2D CNNs (e.g., Xception/EfficientNet/ResNet backbones), followed by simple pooling to reach clip-level decisions, while more recent pipelines incorporate temporal dynamics and semantic motion to improve OoD robustness [2]. Solutions based on visual speech/lip motion have been encouragingly resilient to distribution shifts through the simulation of valid facial motion and articulation [7]. Hybrid CNN-Transformer models—e.g., EfficientNet features fused with Vision Transformers—can achieve competitive accuracy whilst being sensitive to computation and dataset during training and inference [6]. These trends represent a general move away of artifact-specific cues in favor of more generative higher-level temporal semantics, understood to be better inter-generator and post-processing transferable [7].

Perpendicular to appearance and movement, physiological-signal approaches take advantage of rPPG and heartbeat rhythms that are not perfect models of what are already being produced by current generators [8]. DeepRhythm was the first to learn attentional mechanisms to learn heartbeat-related periodicity and did outperform earlier appearance-based baselines on standard benchmarks [8]. The work was inspired by related work that utilized rPPG-based cues not only to distinguish between authenticity but also to assign to the generative source, indicating that biological cues contain residual artifacts that are associated with certain synthesis pipelines [10]. Prior FakeCatcher research also made use of minor changes in blood-volume as identity-coherent cues that cannot be spoofed in space and time, and observed motion, light, and compression sensitivities that drive their combination with more traditional visual properties [15]. Recent surveys argue that detectors still overfit to dataset-specific quirks and degrade on unseen generators and real-platform post-processing, calling for standardized protocols, subject-disjoint splits, and explicit cross-dataset testing, along with evaluation beyond Receiver Operating Characteristic (ROC)/AUC (e.g., calibration, threshold transfer, and reliability analysis) to support deployment-ready systems rather than leaderboard-optimized ones [1].

Recent studies have also explored the application of deep learning for misinformation and deepfake detection in both textual and visual domains. A deep learning-based fake news detection system was proposed using natural language processing techniques, where a Long Short-Term Memory (LSTM) model achieved high classification accuracy by leveraging linguistic and semantic features such as sentiment polarity and readability indices [16]. This work highlights the effectiveness of Artificial Intelligence (AI)-driven approaches in identifying misleading textual content.

In the context of video deepfake detection, Nassri *et al*. [17] presented a ResNet-50-based approach for detecting manipulated videos in both Arabic and English environments. While the model demonstrated strong frame-level performance, the study emphasized critical challenges such as limited dataset diversity, potential overfitting, and the need for subject-disjoint evaluation and cross-dataset validation. These findings further reinforce the importance of robust evaluation protocols and motivate the design of lightweight, reproducible baselines as proposed in this work.

III. Methodology

The methodology consists of four main components: data collection, preprocessing, model design, and experimental evaluation. Fig. 1 illustrates the proposed frame-to-clip deepfake detection pipeline, where sampled

frames are classified using a CNN and aggregated into a clip-level decision.

Each input video is uniformly sampled into a fixed number of frames, which are preprocessed and classified using a ResNet-18 CNN. Frame-level predictions are aggregated using probability averaging (late fusion) to produce a final clip-level decision (real or fake).

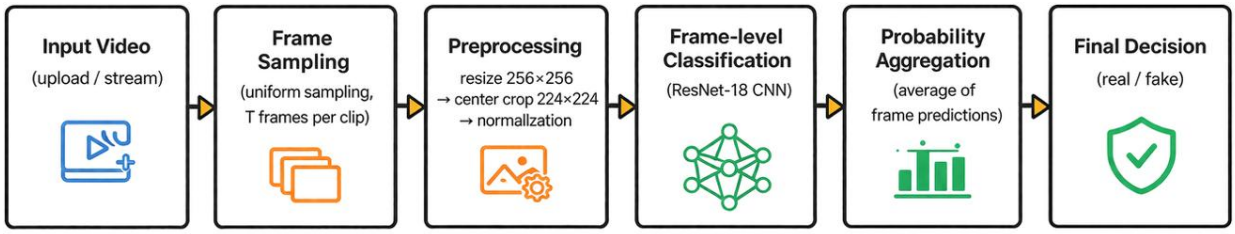


Fig. 1. Overview of the proposed frame-to-clip deepfake detection pipeline.

A. Data Collection and Preparation

1) Data collection

The publicly accessible PolyGlotFake dataset, which is kept in the official Google Drive repository (https://drive.google.com/uc?id=1aBWLii-TbrpKNLSTwpmjqu98eKovWLxF), is used. In our paper, we apply the original public data release and limit our experiments to English (EN) and Arabic (AR) subsets to represent the situation of bilingual deployment.

The data in Table I is the result of limiting PolyGlotFake to the English and Arabic subsets. Identity counts are computed at the subject level (unique speakers/subjects), consistent with the person-disjoint split protocol used in evaluation.

TABLE I. POLYGLOTFAKE EN/AR DATASET COMPOSITION: IDENTITIES AND CLIPS AND LABEL

Language Subset	Label	Unique Identities	Total Clips
EN	fake_EN	695	1328
EN	real_EN	128	128
EN total	-	-	1456
AR	fake_AR	638	2399
AR	real_AR	71	71
AR total	-	-	2470
EN + AR total	-	-	3926

All experiments were conducted using this public dataset snapshot without modification, except for selecting the English and Arabic subsets and organizing them into the directory structure described above.

To prevent identity leakage, we perform subject-disjoint splitting at the person level (no overlap of identities across splits). Unless otherwise stated, the partition is:

(1) Training: 70% (fit model parameters);
(2) Validation: 10% (hyperparameter tuning/early stopping);
(3) Test: 20% (final evaluation).

2) Preprocessing (video → frames)

Identity separation was implemented using the identity identifiers provided in the dataset folder structure and metadata. All video clips belonging to the same identity were grouped together and assigned to a single partition before frame extraction and preprocessing. The assignment was performed programmatically to ensure that no subject appears in more than one subset. To maintain distribution consistency, identities from both language subsets and all labels were proportionally distributed across the training, validation, and test sets so that each split contains samples from both English and Arabic videos as well as real and manipulated clips.

Each video is decoded to RGB and uniformly sampled to a fixed frame budget per clip (e.g., 16–32 frames, depending on duration) so that compute scales predictably while retaining broad temporal coverage. Frames are resized to 256 × 256, center-cropped to 224 × 224, and ImageNet-normalized to match ResNet-18 pretraining. To keep the baseline simple and hardware-friendly, we do not apply face detection/cropping; optional face-tracking crops can be explored in ablations.

B. Model Architecture

We adopt a frame-level ResNet-18 classifier (ImageNet initialization). The final fully connected layer is replaced with a 2-way head (real vs fake), followed by SoftMax. For a clip with T sampled frames and per-frame SoftMax $p_t \in [0,1]^2$, the clip score is computed via probability averaging (late fusion):

$$\bar{p} = \frac{1}{T}\sum_{t=1}^{T} P_i, \quad \hat{y} = \arg\max \bar{p} \tag{1}$$

This frame-to-clip paradigm stabilizes predictions against transient artifacts without requiring heavier temporal models.

C. Video Processing Integration

The vision pipeline converts raw videos → normalized frames → CNN posteriors → clip decision:

(1) Decode & sample: uniformly sample T frames per clip.
(2) Spatial ops: resize → center crop (224) → ImageNet normalization.
(3) Infer: ResNet-18 produces per-frame class probabilities.
(4) Fuse: average per-frame probabilities to a single clip-level score.

(5) Decide: assign the clip label from the fused score.

Light, class-symmetric augmentations (random horizontal flip, mild scale jitter, light JPEG re-encoding) are used only on training frames to improve generalization while preserving forgery cues.

D. Implementation Details

The given model is based on a pretrained ResNet-18 network (ImageNet state) fine-tuned on binary classification (real vs. fake) by substituting the last fully connected layer with a two-class output and SoftMax activation. Categorical cross-entropy loss and stochastic gradient descent with momentum are used to perform training and early stopping is implemented based on validation performance to avoid overfitting. Uniform frame sampling is used to process videos, and predictions at the clip level are done by averaging per-frame class probabilities. In this design, computational efficiency and reproducibility are placed above the architectural complexity.

Stochastic Gradient Descent with Momentum (SGDM) was used to train the network with a momentum coefficient of 0.9 and an initial learning rate of 1×10^{-3}. The learning rate was based on a step decay schedule, which decreased by 0.1 after every 10 epochs. The mini-batch size was 32 and the maximum number of epochs was 30 and early stopping was done based on validation loss to avoid overfitting. Training frames were only augmented with data but random horizontal flipping probability = 0.5, mild scale jitter (±10%), and JPEG re-encoding with quality factors of 7095 was used to enhance the resistance to compression artifacts. Validation and test frames were not augmented (see Table II).

TABLE II. TRAINING CONFIGURATION

Parameter	Value
Optimizer	SGD with momentum
Momentum	0.9
Initial learning rate	$1e^{-3}$
LR schedule	Step decay (× 0.1 every 10 epochs)
Batch size	32
Epochs	30
Early stopping	Validation loss
Augmentation	Flip, scale jitter, JPEG

E. Experimental Setup

All the experiments are performed with the Deep Learning Toolbox of MATLAB. The validation performance is used to select the model, and the best checkpoint is used during the final evaluation. MATLAB and toolbox version, random seed, and decoding stack (FFmpeg/OpenCV) are stored so that they can be reproduced. Assessment is done on a clip basis, and the analysis code calculates the entire metric suite-accuracy, sensitivity/recall, specificity, precision, and Negative Predictive Value (NPV), F1-Score, Matthews Correlation Coefficient (MCC), and macro-AUC and produces confusion matrices, ROC curves, and training/validation plots.

F. Evaluation Metrics

We give a report on clip-level performance on the held-out test identities with:

(1) Accuracy: overall correctness.
(2) Sensitivity/recall and specificity: per-class detection hit and true-negative rates (macro-averaged).
(3) Precision and NPV.
(4) F1-Score: harmonic mean of precision and recall.
(5) MCC: balanced correlation of predictions vs. truth.
(6) AUC (macro, one-vs-all) and ROC curves per class.
(7) Confusion matrix: normalized row/column summaries to inspect error structure.

IV. RESULTS AND DISCUSSION

A. Real-World Applicability and Illustrative Use Cases

Lightweight deepfake detection systems are increasingly applied in real-world scenarios. For example, large-scale platforms such as Facebook (Meta) have actively invested in deepfake detection as part of their media integrity efforts, organizing the DFDC to advance automated detection technologies and support large-scale identification of manipulated media [3]. In digital journalism, organizations such as Reuters and the British Broadcasting Corporation utilize AI-based media verification tools to assess the authenticity of user-generated content, particularly in time-sensitive news reporting [18]. Furthermore, deepfake detection techniques are being integrated into biometric authentication systems to prevent spoofing attacks, where combining face recognition with liveness detection improves the reliability of identity verification against synthetic media [10]. These real-world applications demonstrate the practical relevance and deployment potential of efficient deepfake detection frameworks.

B. Overall Performance

Using the frame-to-clip ResNet-18 baseline, the system achieves an accuracy of 0.9775, sensitivity/recall of 0.9775, specificity of 0.9925, precision of 0.9777, Negative Predictive Value (NPV) of 0.9925, F1-Score of 0.9776, Matthews Correlation Coefficient (MCC) of 0.9701, and macro-average area under the receiver operating characteristic curve using the one-vs-all approach [macro-AUC (OvA)] = 0.9993 on the held-out test identities, as reported in Table III.

TABLE III. CLIP-LEVEL TEST PERFORMANCE ON POLYGLOTFAKE (EN/AR) USING RESNET-18

Metric	Value
Accuracy	0.9775
Sensitivity (Recall)	0.9775
Specificity	0.9925
Precision	0.9777
NPV	0.9925
F1-Score	0.9776
Matthews Correlation Coefficient (MCC)	0.9701
Macro-AUC (OvA)	0.9993

These outcomes show a high level of separability between real and fake clips with an extremely low false-positive rate and a balanced level of precision and

recall. With a person-disjoint evaluation protocol, frame-level CNN inference with late-fusion averaging provides consistent and stable clip-level predictions of unseen identities. On the whole, the suggested implementation is a good performance-efficiency trade-off, which is highly accurate without explicit temporal modeling, and thus can be effectively used in the baseline comparison and practical implementation.

C. Training Dynamics

The training validation curves show that the accuracy grows very fast at the initial training stages and reaches a point of over 95% and the loss continues to decline without any indication of disastrous overfitting (Fig. 2). The stochastic mini-batch optimization will lead to minor oscillations around the minimum, and the last checkpoint chosen through an early-stopping procedure on the validation split is likely to be able to generalize to unseen test identities.

In Fig. 2, the curves exhibit fast convergence and stabilizing optimization and no signs of overfitting because of the early termination criterion of validation performance.

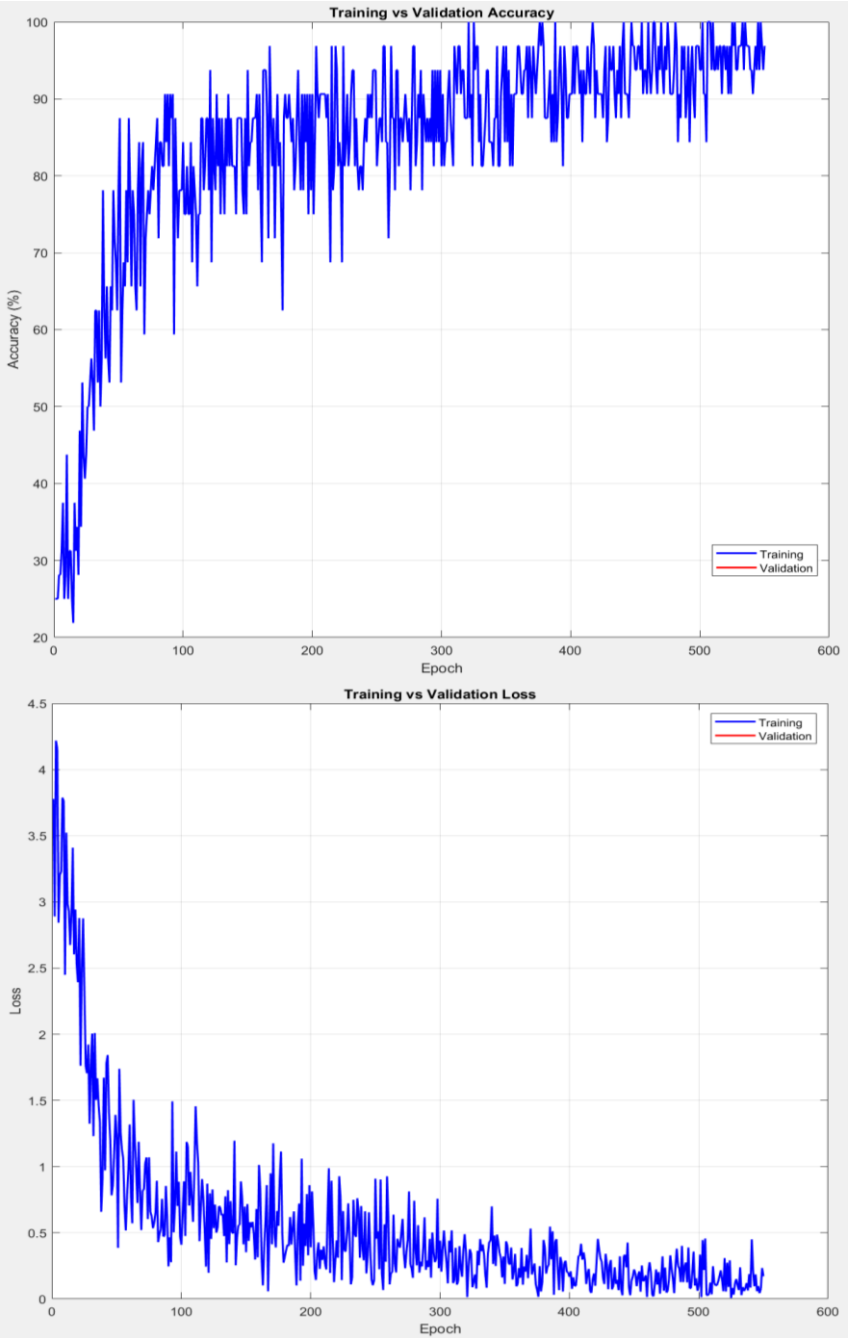


Fig. 2. Training and validation accuracy and loss curve of the ResNet-18 frame-level classifier.

D. Confusion Structure and Per-Class Behavior

The analysis of confusion-matrix (row/column-normalized; see Fig. 3) demonstrates that in this run there are four labels, AR, EN, to AR and to EN, and the recall

is 97% in case of AR, 100% in case of EN, 100% in case of to AR, and 94% in case of to EN, that is, the few faulty results are concentrated on a limited number of cases of to EN. The high purity of the predicted-class (at least 96,100%) expressed in column-wise summaries is in line with the high precision which was mentioned above. These trends are consistent with both the F1 and MCC on the whole and prove that there are no mistakes and no one class prevails.

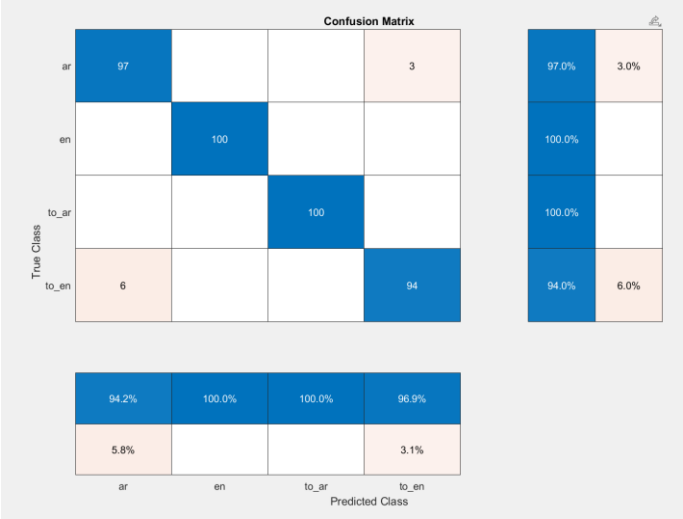


Fig. 3. Normalized confusion matrix at the row and column levels of clip-level classification on the PolyGlotFake (EN / AR) test set. Large diagonal values show that there is a high level of class-wise recall and precision and the error is confined to a limited number of samples.

E. ROC Characteristics and Operating Points

ROC curves per-class (Fig. 4) indicate that AUC = 1.00 across all labels, which agrees with the macro-AUC = 0.9993, indicating almost perfect separability. This facilitates threshold flexibility: a high-specificity point should be selected to allow low-risk deployment (to reduce false accusations); a high-recall point should be used to reduce false discoveries (when doing discovery/triage).

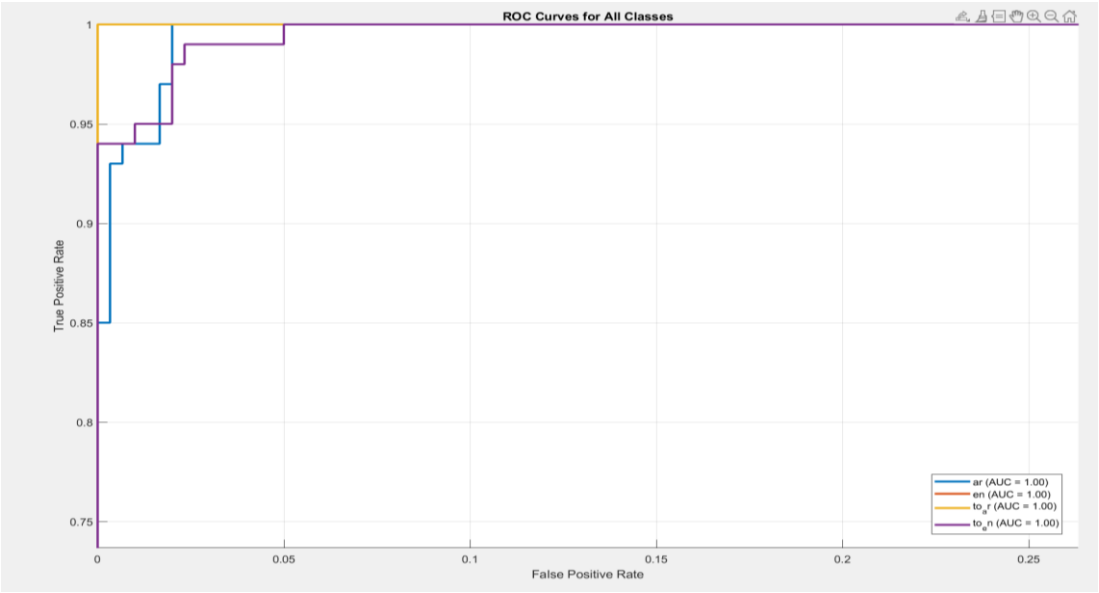


Fig. 4. ROC curves of all classes and AUC. The AUC scores are close to unity, which means that the separability of real and fake classes is good regardless of the operating threshold.

F. Drives Performance

The two design options probably play the greatest role:

(1) Frame-to-clip averaging smooths transient artifacts (motion blur, compression spikes), stabilizing decisions across frames and improving both precision and recall.

(2) ImageNet-aligned preprocessing (resize → center-crop 224; normalization) yields low-variance inputs compatible with the pretrained backbone, aiding convergence without heavy temporal modules.

G. Comparative Perspective and Practical Challenges

This paper aims at getting a powerful and repeatable baseline instead of the extended comparison of different architectures. ResNet-18 has been chosen because of its good trade-off between representational capacity and computational cost, and thus it is well utilized in a delegation on both baseline evaluation and deployment-constrained settings. ResNet-18 is less costly to train and infer than its deeper CNNs or hybrid CNN-Transformer counterparts in the literature and still can be competitive on the PolyGlotFake (EN/AR) subsets.

In experimentation, a number of feasibility difficulties were witnessed. First, frame-level CNN posteriors may be vulnerable to the appearance of transient visual artifacts due to compression, motion blur or changes in lighting. Frame-to-clip probability averaging was observed to reduce this instability by averaging per-frame variations and enhancing the uniformity of decisions at the clip level. Second, there is no explicit temporal modeling, hindering sensitivity to fine motion-based clues, possibly the reason behind the small number of errors on particular subsets (e.g., to EN clips). Lastly, the lightweight architecture can reproducibly and efficiently scale to more complex temporal or multimodal methods but might fail to deliver performance in more challenging cross-dataset situations. The extensions that will be inspired by these observations are temporal aggregation and audio-visual fusion.

The suggested model has a high score in the PolyGlotFake (EN/AR) subsets, but the findings must be seen as a dataset-optimal baseline as opposed to an entirely general deepfake detector. Cross-dataset generalization is also a well-known problem of deepfake forensics because the performance may decline when tested on unseen manipulation patterns or datasets. In this, we are concerned with developing a clear and repeatable base in a controlled dataset environment. Future research will thus consist of cross-dataset assessment on popular benchmarks including DFDC, Celeb-DF and WildDeepfake, and systematic ablation experiments to investigate the effects of frame sampling, preprocessing schemes, as well as temporal modeling.

H. Error Analysis and Practical Guidance

(1) The few observed mistakes involve to_en clips, suggesting sensitivity to specific post-processing or content patterns within that subset. In practice:
(2) Tune the threshold on a representative validation set to match your risk profile (maximize F1 or Youden's J, or impose a specificity floor).
(3) Triage near-boundary scores (e.g., 0.45–0.55) to a human reviewer to minimize costly errors.
(4) As an ablation, consider face-aligned crops and a slightly larger frame budget (e.g., 24–32) to test whether tighter spatial focus or more temporal evidence reduces the remaining to_en errors.

I. Reproducibility: MATLAB Code Used for Metrics & Plots

The MATLAB implementation used to compute accuracy, sensitivity, specificity, precision, Negative Predictive Value (NPV), F1-Score, Matthews Correlation Coefficient (MCC), and macro-AUC using the one-vs-all approach [macro-AUC (OvA)], as well as to generate the confusion matrices, ROC curves, and training and validation plots, is provided to ensure full reproducibility of the reported results. All experiments can be reproduced using the saved YTrue, YPred, scores, trainedNet, and info variables.

J. Real-World Applicability and Illustrative Use Cases

Lightweight video deepfake detection models like the one proposed in the frame-to-clip CNN pipeline have already been implemented or tried in a number of actual applications. By way of illustration, frame-based CNN detectors are utilized by social media platforms, and content moderation pipelines as an initial screening mechanism to identify potentially manipulated videos to be viewed by humans. High specificity and low false-positive rates, which the proposed method can ensure, are essential in such environments because they reduce instances of content being removed improperly and reputational damage.

Video forensics tools are becoming a regular part of digital journalism and reality-checking processes that are implemented to identify the authenticity of user-generated content in the course of breaking news. A video credibility detector like the proposed ResNet-18 baseline, which is resource-efficient can be deployed on commodity hardware to quickly evaluate the video credibility before it is published.

Also, identity verification and biometric security systems are being studied in the detection of deepfakes, specifically to secure video-based authentication against spoofing attacks. The person-disjoint assessment tool adopted in the research is realistic in deployment practice where unrecognized identities have to be evaluated. The above illustrations indicate that the proposed solution is in line with the existing operational needs and can be used as a realistic foundation for actual implementation and additional system integration.

V. Conclusion And Future Work

This paper presented a reproducible and computationally efficient baseline for video deepfake detection on PolyGlotFake (EN/AR). The pipeline uniformly samples frames, applies ImageNet-aligned preprocessing, and uses ResNet-18 with late fusion to produce clip-level decisions. The proposed method achieved 97.75% accuracy, a macro-AUC of 0.999, and an MCC of 0.970 on a subject-disjoint test split, demonstrating strong class separability and stable performance across unseen identities (see Table III and Figs. 2–4). Confusion-matrix analysis shows that errors are rare and not concentrated in a single class, while training and validation curves indicate stable optimization without overfitting. These results validate the frame-to-clip approach as a strong, transparent baseline suitable for bilingual scenarios and practical deployment.

The proposed framework is applicable to real-world scenarios such as social media content moderation, digital

journalism verification, and biometric authentication systems, where robustness to unseen identities and low false-positive rates are essential. Future work will focus on extending this baseline through lightweight temporal modeling (e.g., Gated Recurrent Unit (GRU), LSTM, and temporal convolutional networks), evaluating face-aligned and face-tracked inputs, and integrating audio–visual fusion for cross-modal consistency. Additionally, robustness will be assessed via cross-dataset evaluation and post-processing stress tests, while further investigation will explore model calibration, threshold transferability, and human-in-the-loop decision strategies to support reliable real-world deployment.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Rawad Abdulghafor conceptualized the study, designed the research methodology, supervised the overall work, contributed to the implementation of the frame-to-clip CNN baseline, analyzed the results, and wrote the original draft of the manuscript. Yousuf Al Husaini contributed to the research design, validation of the experimental results, discussion of the findings, and critical revision of the manuscript. Mohammad Abrar contributed to the literature review, methodology refinement, data organization, and manuscript editing. Abdullah Said Al-Aamri contributed to the theoretical background, evaluation of the experimental approach, interpretation of the results, and critical review of the manuscript. Mohammed Abdulla Salim Al Husaini contributed to the experimental setup, performance evaluation, results analysis, and technical revision of the manuscript. Wasin AlKishri contributed to the literature review, writing support, validation of the manuscript content, and final editing. All authors reviewed and approved the final version of the manuscript.

FUNDING

This research was funded by the Arab Open University–Oman, and the Ministry of Higher Education, Research, and Innovation, Undergraduate Research Grant (Fund No.: BFP/URG/ICT/24/002).

ACKNOWLEDGMENT

The authors would like to thank all who volunteered for this experiment. We thank our colleagues for the insightful feedback on the earlier iteration of this paper as well. The authors are also grateful to the Arab Open University–Oman for support.

REFERENCES

[1] F. A. Croitoru, N. Ristea, R. T. Ionescu *et al.*, "Deepfake media generation and detection in the generative AI era: A survey and outlook," arXiv Preprint, arXiv:2411.19537, 2024.

[2] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[3] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The DeepFake detection challenge (DFDC) dataset," arXiv Preprint, arXiv:2006.07397, 2020.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207–3216.

[5] B. Zi, H. Chang, J. Ma, L. Jiang, D. Yang, and Y. Liu, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM International Conference on Multimedia*, 2021, pp. 2382–2390.

[6] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *International Conference on Image Analysis and Processing*, 2021, pp. 219–229.

[7] A. Haliassos, J. Ma, S. Zafeiriou, and V. Kumar, "Lips don't lie: A generalizable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5039–5049.

[8] H. Qi, Q. Guo, F. Juefei-Xu *et al.,* "DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020, pp. 4318–4327.

[9] S. M. Qureshi, M. M. Khan, A. R. Javed *et al*., "Deepfake forensics: A survey of digital forensic methods for detection," *PeerJ Comput. Sci.*, vol. 10, e1928, 2024.

[10] U. A. Ciftci, İ. Demir, and L. Yin, "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2020, pp. 1–10.

[11] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2889–2898.

[12] Y. He, H. Xue, Q. Wang *et al*., "ForgeryNet: A versatile benchmark for comprehensive forgery analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4369–4378.

[13] P. Kwon, D. Y. Park, K. J. Yoon, and D. Kim, "KoDF: A large-scale Korean DeepFake detection dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10744–10753.

[14] Z. Yan, Y. Li, H. Qi *et al*., "DeepfakeBench: A comprehensive benchmark of deepfake detection," arXiv Preprint, arXiv:2307.01426, 2023.

[15] U. A. Ciftci and İ. Demir, "FakeCatcher: Detection of synthetic portrait videos using biological signals," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. doi: 10.1109/TPAMI.2020.3009287

[16] M. S. Al-Hattali, R. Abdulghafor, W. AlKishri, Y. Al Husaini, M. A. S. Al Husaini, and M. Abrar, "Fake news detection via AI model," in *Proc. Int. Conf. Comput. Appl. (ICCA)*, Bahrain, 2025, pp. 1–7. doi: 10.1109/ICCA66035.2025.11430952

[17] R. S. Al Nassri, R. Abdulghafor, M. A. Salim Al Husaini, W. AlKishri, Y. A. Husaini, and M. Abrar, "Video deepfake detection in English and Arabic," in *Proc. Int. Conf. Comput. Appl. (ICCA)*, Bahrain, 2025, pp. 1–6. doi: 10.1109/ICCA66035.2025.11430955

[18] T. M. Mitchell *et al*., "Media forensics and deepfake detection in journalism," *IEEE Computer*, vol. 54, no. 3, pp. 38–45, 2021.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).