KMANPUB | AI and Tech in Behavioral and Social Sciences Vol. 5 No. 1 (2027): 1-14

Article history:
Received 24 April 2026
Revised 14 July 2026
Accepted 24 July 2026
Initial Publication 31 July 2026
Final Publication 01 January 2027

Compression-Resilient Deepfake Detection for AI-Assisted Social Media Content Monitoring: Feature-Level Fusion of ResNet50 and EfficientNet-B0 with a Learnable Visibility Matrix

Samir Qaisar Ajmi[1], Mahmood Deypir[1*], Razieh Farazkish[1], Ali Broumandnia[1]

[1] Department of Computer Engineering, ST.C., Islamic Azad University, Tehran, Iran

*** Corresponding author email address**: mdeypir@iau.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Qaisar Ajmi, S., Deypir, M., Farazkish, R., & Broumandnia, A. (2027). Compression-Resilient Deepfake Detection for AI-Assisted Social Media Content Monitoring: Feature-Level Fusion of ResNet50 and EfficientNet-B0 with a Learnable Visibility Matrix. *AI and Tech in Behavioral and Social Sciences*, *5*(1), 1-14.
https://doi.org/10.61838/kman.aitech.5856

A B S T R A C T

Deepfakes circulating through social media create a combined technical, behavioral, and governance challenge because manipulated videos may be redistributed after resizing, re-encoding, and compression. This study evaluated a frame-level detector combining ResNet50 and EfficientNet-B0 through feature-level fusion and an iterative Visibility Matrix procedure. Frames were obtained from FaceForensics++ c23 and Celeb-DF v2 and resized to 128 × 128 pixels. The fused model produced the highest reported point estimates, including 83% validation accuracy, an AUC-ROC of 0.89, balanced accuracy of 0.81, and approximately 18 ms inference time per image. Reported accuracy was approximately 79% on an additional unseen evaluation and approximately 78–80% under a severe-compression condition, compared with approximately 65–67% for ResNet50. These results are descriptive because the available experiment did not report repeated runs, confidence intervals, formal statistical comparisons, or a fully specified external-testing protocol. The detector should therefore be interpreted as a candidate frame-level triage component requiring video-disjoint validation, formal ablation, calibrated thresholds, and human review before operational use.

Keywords: *deepfake detection; social media content monitoring; content moderation; misinformation; digital trust; compression robustness; ResNet50; EfficientNet-B0; human-in-the-loop AI*

1. Introduction

Synthetic facial media have progressed from visibly imperfect demonstrations to highly realistic images and videos capable of imitating identity, expression, and speech. Although this technology has legitimate uses in entertainment, education, accessibility, and creative production, deceptive applications can support impersonation, fraud, non-consensual imagery, political manipulation, and coordinated disinformation. The problem is especially consequential on social media, where audiovisual content can be copied, reframed, and redistributed at high speed. Experimental evidence suggests that political deepfakes may generate uncertainty and reduce trust in news encountered through social platforms, even when viewers are not fully deceived (Vaccari &

Chadwick, 2020). Legal and policy scholarship has similarly emphasized risks to privacy, democratic discourse, reputation, and national security (Chesney & Citron, 2019). Deepfake detection is therefore not only a computer-vision problem; it is also part of the broader socio-technical challenge of maintaining trustworthy online communication.

The operational monitoring problem is substantially harder than benchmark classification. Videos distributed through social networks and messaging services are commonly resized, cropped, re-encoded, and heavily compressed. These operations suppress high-frequency artifacts and subtle blending boundaries on which many detectors depend. At the same time, content-monitoring systems must process large volumes of heterogeneous material, support rapid prioritization, and avoid false accusations against legitimate users. Human observers and machine detectors also make different kinds of errors, and human-machine collaboration can outperform either alone; however, inaccurate model advice may cause reviewers to revise correct judgments in the wrong direction (Groh et al., 2022). Consequently, an effective platform-oriented detector should function as a calibrated decision-support tool within a human-in-the-loop workflow rather than as an autonomous arbiter of truth.

The behavioral relevance of the problem is reinforced by the increasing perceptual credibility of synthetic faces. AI-generated faces have been shown to be difficult to distinguish from real faces and, under some conditions, may even be judged as more trustworthy (Nightingale & Farid, 2022). Such perceptual effects can increase the persuasive potential of manipulated content and make simple user vigilance an inadequate safeguard. Social media monitoring therefore requires a combination of automated forensic screening, contextual assessment, provenance information, media literacy, and accountable platform procedures. The objective is not to classify every unusual video as harmful, but to identify material that warrants closer review before it contributes to misinformation cascades, identity harm, or erosion of confidence in authentic media.

Convolutional neural networks remain strong feature extractors for facial-forensics tasks. ResNet50 uses residual connections to support optimization of a deep architecture and is effective at capturing local texture, edges, and morphological irregularities (He et al., 2016). EfficientNet-B0 uses compound scaling and mobile inverted bottleneck blocks to balance accuracy, parameter count, and computational cost (Tan & Le, 2019). Its squeeze-and-excitation mechanisms adaptively recalibrate channel responses and can support sensitivity to broad structural and illumination patterns (Hu et al., 2018). Because these architectures emphasize different visual properties, feature-level fusion may produce a richer representation than either model alone.

Recent research has increasingly addressed low-quality and cross-quality deepfake detection. Spatial restoration has been used to recover manipulation evidence in degraded social-media videos (Li et al., 2023). Frequency-aware attention and multi-view knowledge distillation have been proposed for compressed images [11]. Quality-agnostic collaborative learning (Le & Woo, 2023), multi-scale branch zooming (Lee et al., 2022), high-frequency enhancement (Gao et al., 2024), learnable spatial-rich-model filtering (Han et al., 2021), and knowledge-distillation strategies (Kim et al., 2021) all seek to reduce dependence on a single manipulation quality or dataset. Attention-based (H. Zhao et al., 2021), identity-aware (Cozzolino et al., 2021), self-consistency (T. Zhao et al., 2021), geometric (Sun et al., 2021), and feature-restoration methods [21] further show that robust detection requires cues that survive changes in quality and generation technique.

The present study addresses this problem through a hybrid frame-level architecture designed with social media content monitoring in mind. ResNet50 and EfficientNet-B0 were separately adapted to deepfake data and then combined at the feature level. A learnable Visibility Matrix was used to expose weak manipulation traces under compression and illumination changes. The technical objective was to improve discrimination under degraded conditions; the applied objective was to examine whether the resulting speed and robustness could support automated pre-screening of platform video. The study does not treat the classifier score as a final determination of authenticity. Instead, it positions the detector as one component of a layered monitoring process that can prioritize suspicious media for human review, provenance verification, and additional multimodal analysis.

2. Related Work

Benchmark design has strongly influenced the development of deepfake detection. FaceForensics++ includes real videos and several face-manipulation methods under raw and compressed conditions, making it useful for

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

studying the loss of forensic evidence during re-encoding (Rössler et al., 2019). Celeb-DF v2 contains more visually convincing identity swaps and is commonly used as a cross-dataset generalization benchmark (Dolhansky et al., 2020). The diversity of these datasets is important because deep detectors may learn dataset-specific acquisition or compression signatures instead of manipulation mechanisms. Cross-dataset evaluation is therefore a more demanding indicator of practical robustness than in-domain accuracy alone (Mirsky & Lee, 2021; Tolosana et al., 2020; Verdoliva, 2020).

A major line of work focuses on spatial and texture traces. Residual filters, edge inconsistencies, local blending boundaries, and high-frequency artifacts can reveal synthesis even when the image appears plausible to human observers. Han et al. combined two streams with a learnable spatial-rich-model component to capture manipulation residuals (Han et al., 2021). Gao et al. emphasized high-frequency information for highly compressed content (Gao et al., 2024), while Wang et al. applied frequency-domain filtered residual learning (Wang et al., 2023). Texture-based and feature-restoration approaches also attempt to preserve or recover weak signals after compression (Ke & Wang, 2023; Yu et al., 2021). These methods support the use of ResNet50 as a local and residual feature extractor in the current architecture.

A second line of work uses global, multi-scale, or semantic consistency. Multi-attentional detectors allocate attention to several facial regions rather than depending on a single artifact (H. Zhao et al., 2021). Multi-scale convolution combined with vision transformers can integrate local and long-range evidence (Lin et al., 2023). Identity-aware systems analyze whether facial motion and appearance remain compatible with the claimed person [18], while self-consistency methods examine relationships among different image regions (T. Zhao et al., 2021). EfficientNet-B0 is relevant in this context because its compound-scaled design and channel recalibration provide an efficient way to model distributed image structure (Hu et al., 2018; Tan & Le, 2019).

Compression-robust methods increasingly combine restoration, augmentation, distillation, and quality adaptation. Li et al. proposed a spatial restored detection framework for low-quality social-network video (Li et al., 2023). Woo introduced frequency attention with multi-view distillation for compressed images (Woo, 2022). Le and Woo developed intra-model collaborative learning across quality conditions (Le & Woo, 2023), and Lee et al. proposed unsupervised multi-scale branch zooming for low-quality videos (Lee et al., 2022). Super-resolution preprocessing has also been studied as a way to improve low-resolution detection (Perera et al., 2022). These studies show that robustness cannot be obtained solely by increasing model depth; the training process must explicitly represent degradation.

Ensemble and fusion strategies are attractive because no single representation captures every manipulation trace. Late fusion combines classifier probabilities, whereas feature-level fusion integrates hidden representations before classification. The latter preserves more information but may increase dimensionality and computational demand. The current model concatenates pooled ResNet50 and EfficientNet-B0 vectors before a compact fully connected head. Grad-CAM was used qualitatively to examine whether the fused detector attended to plausible facial regions (Selvaraju et al., 2017). Such interpretability is useful because deepfake detectors may otherwise rely on unstable shortcuts or background artifacts.

A separate but complementary literature examines deepfake detection as a problem of social media governance and human decision making. Research on online misinformation emphasizes that technical detection must be integrated with behavioral, institutional, and communication-based interventions (Lazer et al., 2018). Human-machine experiments show that combined judgments may improve accuracy, but they also demonstrate the risk of automation bias when an incorrect model prediction influences a human reviewer (Groh et al., 2022). Content moderation scholarship further shows that platform decisions are socially consequential because ranking, labeling, demotion, and removal practices shape visibility, participation, and perceptions of legitimacy (Gillespie, 2018). These findings motivate a monitoring framework in which algorithmic outputs are calibrated, interpretable, contestable, and embedded within transparent review procedures.

3. Methods and Materials

3.1. Study Design and Datasets

This computational experimental study developed and compared three frame-level binary classifiers: ResNet50, EfficientNet-B0, and a feature-fusion model combining both backbones. The workflow included dataset preparation, transfer learning, visibility-guided adaptation, feature fusion, and descriptive evaluation under standard,

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

additional unseen, and compressed conditions. The primary endpoint was frame-level classification performance. The social-media monitoring workflow discussed later is a proposed application context and was not evaluated as a complete platform-level system.

Two public benchmarks were used. FaceForensics++ c23 contains moderately compressed real and manipulated videos generated using DeepFakes, Face2Face, FaceSwap, and NeuralTextures (Rössler et al., 2019). The available study record reports approximately 890 real and 1,000 manipulated videos and ten sampled frames per video, yielding approximately 19,000 images. Celeb-DF v2 contains approximately 900 real and 5,600 manipulated videos (Li et al., 2020), from which ten frames per video were reportedly sampled. Exact inclusion, exclusion, failed-processing, class, and split counts were not available in the source record; these quantities should be reported in any reproducible implementation.

Table 1

Dataset and preprocessing summary

Dataset	Real videos	Fake videos	Frames per video	Approx. extracted frames	Primary role
FaceForensics++ c23	≈890	≈1,000	10	≈19,000	Training and compressed benchmark material
Celeb-DF v2	≈900	≈5,600	10	>65,000	Higher-quality and unseen/generalization material

3.2. *Frame Preparation*

Frames were sampled at regular temporal intervals, resized to 128 × 128 pixels, normalized, and organized by binary class. Adaptive class weights were applied to the cross-entropy objective, and mixed-precision training was used in a CUDA-enabled environment. The source record did not provide the decoder, face detector, alignment procedure, handling of multiple or undetected faces, normalization constants, or exact augmentation settings. These implementation details are required for exact replication.

The available study record describes separate training and validation sets and an additional unseen evaluation. It does not confirm whether splitting was conducted at the video level before frame extraction or whether identities were disjoint across partitions. Because frames from the same source video are highly correlated, the reported metrics should be considered provisional until strict video-disjoint and preferably identity-disjoint splitting is documented.

3.3. *ResNet50 Baseline*

ResNet50 was initialized with ImageNet-pretrained weights (He et al., 2016). The original classification layer was replaced with a binary head for real-versus-fake prediction. Residual blocks enabled the network to learn local texture, boundary, and edge irregularities while maintaining stable gradient flow. Fine-tuning was performed for 12 epochs using AdamW optimization, a learning rate of 1×10^{-4}, mini-batches of 64 images, dropout of 0.20, class-weighted cross-entropy, and mixed precision. Training loss decreased from approximately 0.47 to 0.01, while validation accuracy stabilized around 74–76%.

The baseline was used both as an independent detector and as the local-feature branch of the final model. Its global-average-pooling layer produced a 2,048-dimensional feature vector.

3.4. *Visibility Matrix Learning*

A learnable Visibility Matrix was introduced to modify image visibility during alternating transformation-learning and detector-adaptation stages. Four stages were reported, each containing three visibility-learning iterations followed by network adaptation. However, the available study record does not provide the matrix dimensions, initialization, transformation equation, hinge-loss formulation, perturbation constraints, optimizer settings, or stopping rule. Consequently, the procedure is not yet independently reproducible and should be accompanied by a formal algorithm or pseudocode.

The Visibility Matrix was intended as a robustness mechanism rather than a conventional enhancement filter. Intermediate stages reduced validation performance, indicating that the transformation could suppress both nuisance information and useful discriminative evidence. Because no controlled fusion-without-matrix ablation was reported, the independent contribution of the Visibility

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Matrix cannot be distinguished from the effects of feature fusion, parameter count, or training exposure.

3.5. *EfficientNet-B0 Branch*

EfficientNet-B0 was initialized with ImageNet-pretrained weights (Tan & Le, 2019). Its mobile inverted bottleneck blocks and squeeze-and-excitation operations provided a lightweight representation of distributed structural and illumination cues (Hu et al., 2018). The original output layer was replaced with a two-class head. Final layers were fine-tuned for ten epochs with AdamW, cross-entropy loss, a learning rate of 1×10^{-4}, normalized 128×128 inputs, and mixed-precision computation. The adaptive-average-pooled output was a 1,280-dimensional feature vector.

The EfficientNet branch was selected to complement ResNet50 rather than replace it. Activation inspection indicated greater sensitivity to broader illumination, symmetry, and face-level structure, whereas ResNet50 placed more emphasis on texture boundaries and local anomalies.

3.6. *Feature-Level Fusion and Classification*

The final architecture removed the original classification heads from both networks and concatenated their pooled feature vectors. The resulting 3,328-dimensional representation was passed to a fully connected layer with 512 units, rectified-linear activation, and dropout of 0.30. A final linear layer produced the two class scores. During the reported fusion training, the internal backbone weights were frozen and the fusion head was optimized for six epochs with Adam, a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, class-weighted cross-entropy, and mixed precision.

The design is a feature-level fusion model rather than a probability-level ensemble. A secondary late-fusion experiment based on averaging classifier probabilities reportedly achieved approximately 79% accuracy, whereas feature-level fusion achieved 83%. Because the models differ in representation dimensionality and trainable parameter count, this comparison should be interpreted descriptively rather than as proof that feature-level fusion is intrinsically superior.

Table 2

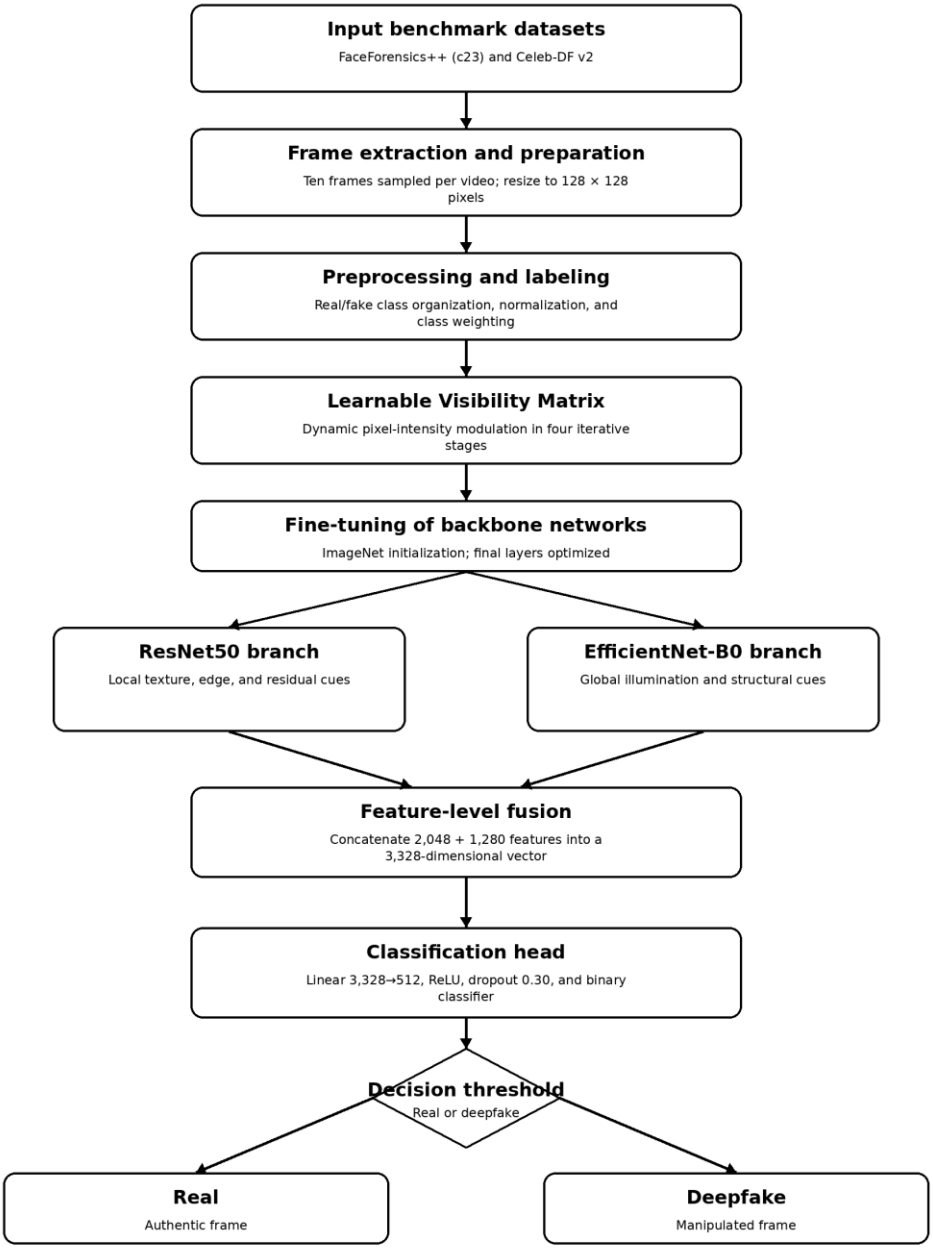
Backbone and fusion configuration

Component	Depth / parameters	Feature output	Principal representation	Trainable stage in fusion
ResNet50	50 layers; ≈25 M	2,048	Local texture, residual, and edge cues	Fusion head; backbone frozen
EfficientNet-B0	≈82 MBConv operations; ≈5.3 M	1,280	Global structure and illumination cues	Fusion head; backbone frozen
Concatenation + MLP	3,328→512; dropout 0.30	512	Joint local-global representation	Trainable
Binary classifier	512→2	2 scores	Real versus deepfake	Trainable

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Figure 1

Experimental workflow for compression-resilient deepfake detection and AI-assisted social media content monitoring.



3.7. Evaluation

Performance was summarized using accuracy, AUC-ROC, balanced accuracy, and inference time per image. Additional analyses reported accuracy on an unseen evaluation and under a severe-compression condition described as quality 40. The source record did not specify the codec, compression software, sample count, partition independence, or exact protocol for this condition. The experiment also reported single point estimates rather than repeated-seed means, confidence intervals, or formal statistical comparisons; all model differences are therefore descriptive.

3.8. Intended Social Media Monitoring Workflow

The proposed detector can be interpreted as the visual-forensics layer of a broader platform monitoring pipeline. In such a pipeline, uploaded or rapidly circulating videos would first undergo standard decoding, face detection, and sparse frame sampling. The detector would then assign frame-level risk scores, which should be aggregated at the

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

video level and combined with contextual signals such as provenance metadata, upload history, user reports, duplication patterns, and audio-visual consistency. Content exceeding a calibrated review threshold would be prioritized for specialist examination rather than automatically removed.

This workflow distinguishes detection from moderation. A model may identify manipulation evidence without establishing the intent, harmfulness, legality, or contextual meaning of a video. Final platform action should therefore incorporate human review, source verification, uncertainty estimates, and an appeal mechanism. The present experiment evaluates only the visual detector and does not test a complete platform system; the monitoring workflow is provided to clarify the intended socio-technical application and the boundary of the reported evidence.

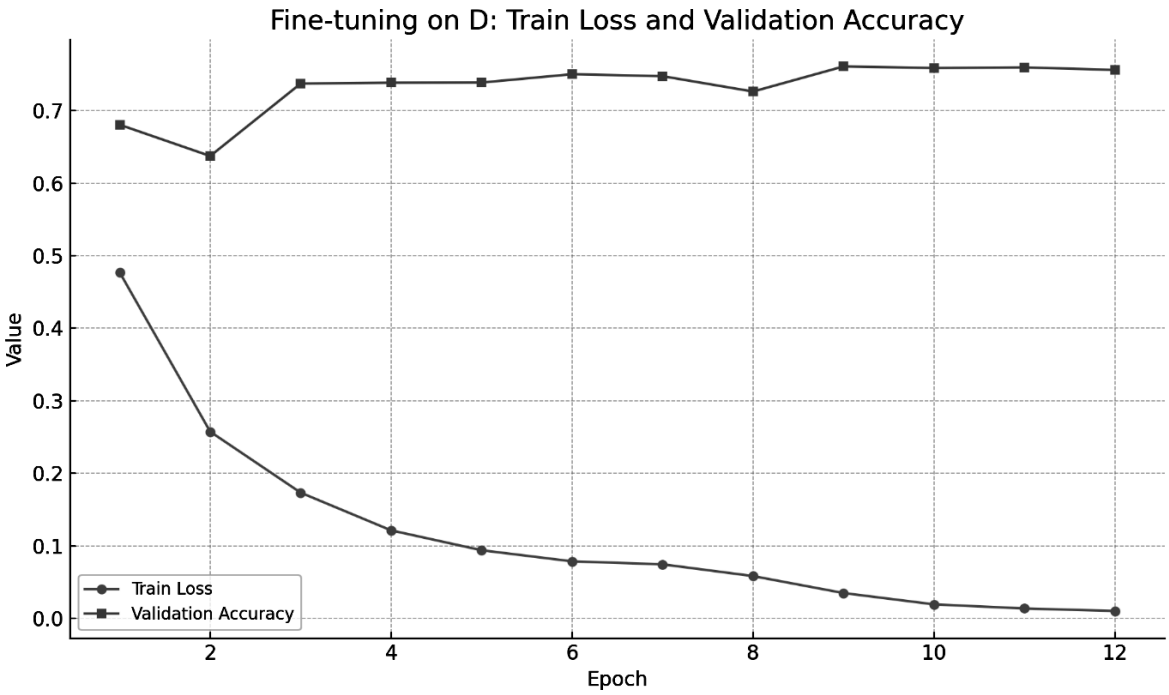
4. Findings and Results

4.1. Baseline Training

ResNet50 converged steadily during 12 fine-tuning epochs. Training loss declined from approximately 0.47 at the beginning of optimization to approximately 0.01 by the final epoch. Validation accuracy fluctuated early and then remained within approximately 74–76%. This pattern indicates successful fitting without a pronounced late-epoch collapse in validation performance. The final reported validation accuracy was 76%, with AUC-ROC of 0.82, balanced accuracy of 0.74, and inference time of approximately 15 ms per image.

Figure 2

ResNet50 fine-tuning: training loss and validation accuracy across 12 epochs.

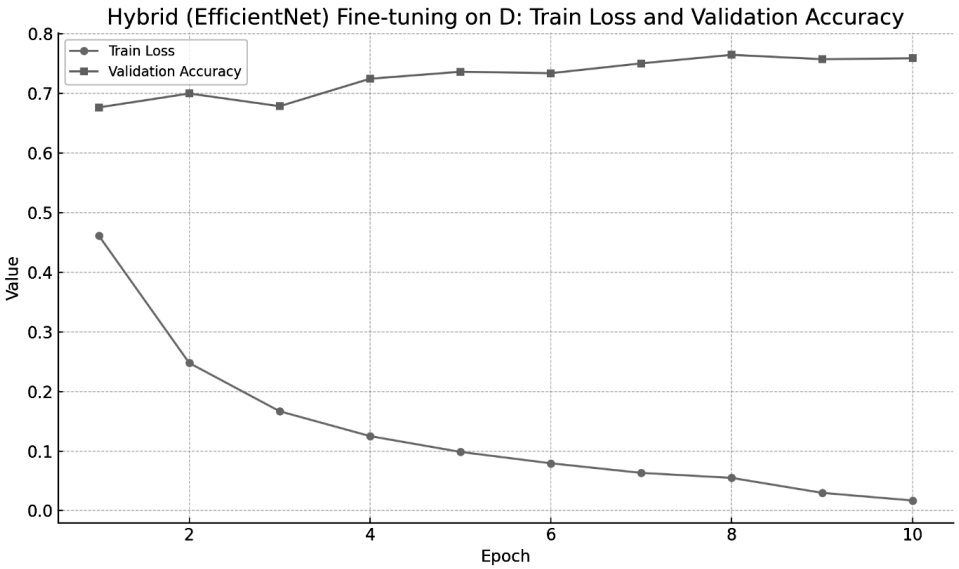


EfficientNet-B0 showed a similar reduction in training loss over ten epochs, from approximately 0.46 to 0.017, while validation accuracy increased from approximately 67% to the 75–76% range during training. Its final comparison accuracy was reported as 74%, with AUC-ROC of 0.80 and inference time of approximately 12 ms per image. Although its accuracy was slightly below that of ResNet50, its lower parameter count and faster inference supported its role as a complementary branch.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Figure 3

EfficientNet-B0 fine-tuning: training loss and validation accuracy across ten epochs.

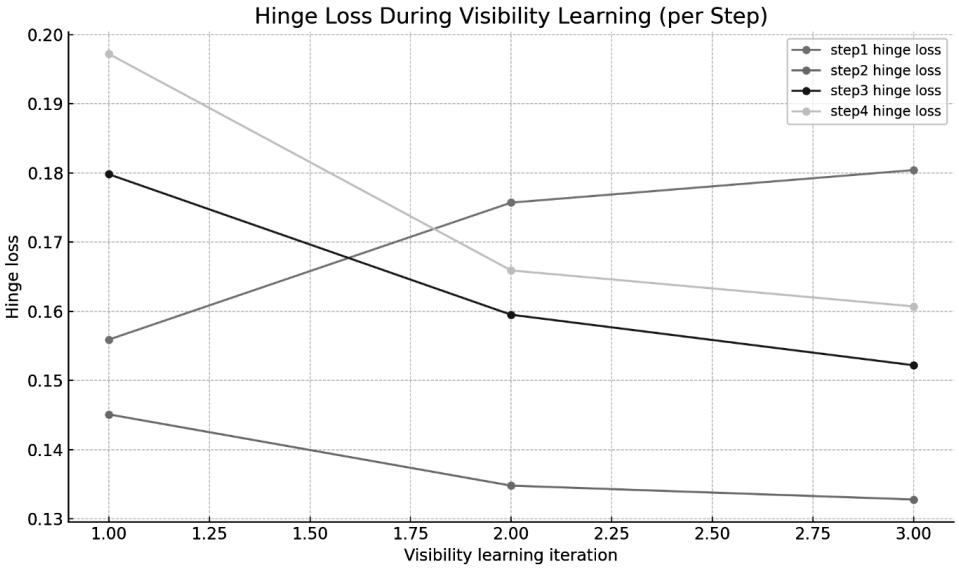


4.2. Visibility-Guided Adaptation

The Visibility Matrix learning curves showed that the transformation did not behave monotonically across stages. In the initial stage, hinge loss increased slightly, consistent with an unstable early transformation. Subsequent stages generally produced declining or stabilizing hinge loss, indicating that the visibility mechanism increasingly identified image regions that challenged the detector in a consistent manner. The original ResNet-based visibility sequence temporarily reduced apparent validation accuracy to approximately 63–69%, but the study interpreted this decrease as the cost of training on more difficult inputs and reported improved robustness on new and degraded samples.

Figure 4

Hinge-loss trajectories during four stages of Visibility Matrix learning with the ResNet50 branch.

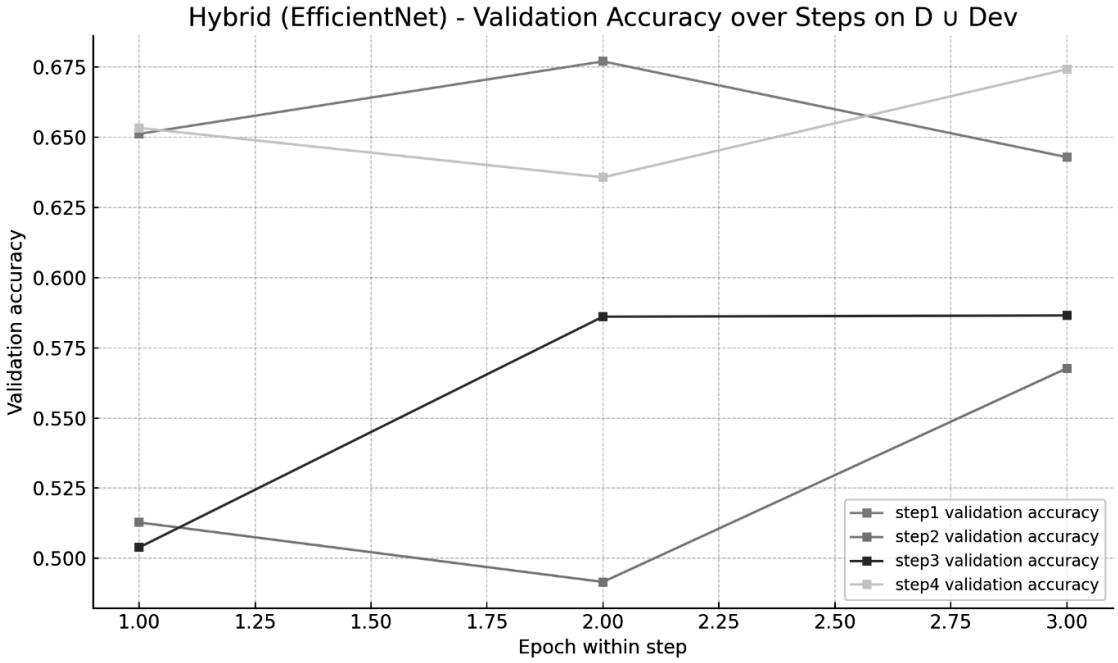


AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

When visibility learning was applied to the EfficientNet-containing hybrid, stage 2 showed the largest validation decline, reaching approximately 49–56%. Stages 3 and 4 recovered, with the final stage reaching approximately 65–67%. The non-monotonic trajectory suggests that overly aggressive input modulation can remove useful evidence, whereas repeated alternation between transformation learning and detector adaptation can restore a more useful balance.

Figure 5

Validation accuracy across Visibility Matrix stages for the EfficientNet-containing hybrid.



4.3. Feature-Fusion Performance

Feature-level fusion produced the highest reported validation point estimates: 83% accuracy, AUC-ROC of 0.89, and balanced accuracy of 0.81, compared with 76%, 0.82, and 0.74 for ResNet50. The reported inference time was 18 ms per image. These differences may reflect complementary representations, additional trainable capacity, or differences in training exposure. A parameter-aware ablation and repeated-run evaluation are needed before attributing the gain specifically to architectural complementarity or the Visibility Matrix.

Table 3

Performance comparison of the evaluated models

Model	Accuracy	AUC-ROC	Balanced accuracy	Inference time	Dominant cue
ResNet50	76%	0.82	0.74	15 ms/image	Local detail and edges
EfficientNet-B0	74%	0.80	Not reported	12 ms/image	Global and illumination patterns
Feature fusion	83%	0.89	0.81	18 ms/image	Complementary local-global features

Grad-CAM inspection suggested attention to facial regions such as the lips, periocular areas, and jaw contours. These visualizations are qualitative and do not demonstrate causal feature use or exclude reliance on dataset-specific backgrounds, compression signatures, or identity cues. Representative correct and incorrect cases across datasets and compression levels, together with quantitative explanation tests, are needed for stronger interpretability claims.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

4.4. *Compression Robustness, Unseen Data, and Monitoring Relevance*

The hybrid detector reportedly retained approximately 79% accuracy on an additional unseen evaluation and approximately 78–80% under a severe-compression condition, whereas ResNet50 reportedly achieved approximately 65–67%. The dataset identity, sample count, class distribution, codec, compression implementation, and model-selection independence of these evaluations were not fully specified. Accordingly, these results should be presented as preliminary robustness observations rather than definitive cross-dataset or deployment evidence. The 18-ms frame-level timing also excludes decoding, face detection, frame sampling, aggregation, and review-system overhead.

Table 4

Robustness and generalization findings relevant to social media content monitoring

Evaluation condition	Reported result	Interpretation
Unseen evaluation data	≈79% hybrid accuracy	Moderate out-of-sample retention
Severe compression (quality 40)	Hybrid ≈78–80%; ResNet50 ≈65–67%	Higher compression resilience
False-positive errors	Up to 30% lower than baseline	Fewer authentic frames labeled fake
Correct fake-video recognition	≈8% higher than baseline	Improved sensitivity to manipulated content
Late fusion versus feature fusion	≈79% versus 83% accuracy	Hidden-feature integration outperformed probability averaging

5. Discussion

The central descriptive result was that feature-level fusion produced higher reported point estimates than the two standalone backbones while adding modest frame-level inference time. The experiment does not isolate whether this gain resulted from complementary representations, increased capacity, the Visibility Matrix, or the training schedule. A complete ablation with identical splits, augmentation, optimization, and repeated seeds is required to identify the source of improvement.

The reported compression results are potentially relevant to deployment, but the compression protocol is insufficiently specified and no common-protocol comparison with established compression-robust methods was performed. The findings therefore do not establish state-of-the-art performance. They indicate only that the fused model retained a reported within-study margin over ResNet50 under the stated condition.

The Visibility Matrix produced an informative but unstable training pattern. Accuracy declines during intermediate stages indicate that robustness transformations can become too aggressive. A transformation that eliminates both nuisance information and discriminative evidence may make the training set harder without improving real-world detection. Recovery in later stages suggests that alternating transformation learning and classifier adaptation can partially correct this problem. Future implementations should constrain perturbation magnitude, monitor perceptual distortion, and use an independent validation set to select the visibility stage. A formal ablation comparing fusion with and without the matrix is also necessary to isolate its contribution.

The increase in AUC-ROC from 0.82 to 0.89 and in balanced accuracy from 0.74 to 0.81 suggests better discrimination in the reported evaluation. However, the absence of repeated runs, confidence intervals, confusion matrices, class-specific metrics, calibration measures, and paired tests prevents assessment of uncertainty. Evaluation should also be conducted at the video level because consecutive frames are correlated and operational decisions are generally made per video rather than per frame.

The reported 18-ms frame inference time describes model inference only and should not be equated with end-to-end high-throughput screening. A reproducible timing protocol should identify the hardware, software versions, batch size, warm-up procedure, number of repetitions, and whether data transfer and preprocessing were included. Platform-level latency additionally includes decoding, face detection, frame selection, aggregation, storage, and human-review capacity.

Cross-dataset generalization remains unresolved. The approximately 79% unseen-data accuracy cannot establish external robustness until the evaluation dataset, identities, manipulation methods, sample counts, and model-selection procedure are fully specified. Strict video-disjoint, identity-disjoint, and generator-disjoint testing on an independently held-out benchmark is required to determine whether the detector learned manipulation evidence rather than dataset signatures.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

5.1. *Implications for AI-Assisted Social Media Content Monitoring*

From an application perspective, the model is most appropriate for content triage rather than autonomous truth determination. A platform could use the detector to identify highly compressed videos that warrant accelerated review, request provenance information, compare reposted versions, or trigger secondary audio-visual and identity-consistency analyses. The model's relatively low frame-level latency is useful in this context because monitoring systems must screen large volumes of material before harmful content becomes widely distributed. Nevertheless, the measured speed is not equivalent to platform-level throughput, and the reported accuracy is not sufficient to justify automatic removal, account restriction, or public labeling without corroborating evidence.

A monitoring system should convert frame-level scores into a transparent video-level risk estimate. Thresholds may be adjusted according to the purpose of review: a lower threshold may be suitable for internal prioritization, whereas a substantially higher evidentiary standard is required for public warnings or enforcement. Risk scores should be combined with provenance standards, metadata inspection, source credibility, user reports, and contextual information. Such a layered approach acknowledges that visual manipulation is only one dimension of harmful content and that an authentic video can still be misleading when selectively edited, falsely captioned, or removed from its original context.

The system may also support behavioral and social research by enabling structured analysis of when, where, and in what form suspected synthetic media circulate. Aggregated and privacy-preserving monitoring data could help researchers examine exposure patterns, recirculation, user responses, correction behavior, and the relationship between synthetic media and trust. Any such use should avoid inferring personal characteristics or political beliefs from users and should separate content-level forensic signals from judgments about individuals or groups.

5.2. *Human-Centered Moderation, Trust, and Governance*

The social value of a detector depends not only on accuracy but also on how its output influences human judgment. Deepfakes may create generalized uncertainty and reduce trust in social media news (Vaccari & Chadwick, 2020), while highly realistic synthetic faces may be perceived as credible (Nightingale & Farid, 2022). At the same time, machine suggestions can improve collective detection yet also induce automation bias when the model is wrong (Groh et al., 2022). Review interfaces should therefore display uncertainty, relevant evidence regions, model limitations, and alternative explanations rather than a categorical "real/fake" verdict. Reviewers should be trained to treat the score as one source of evidence and to document the basis of final decisions.

Responsible deployment also requires procedural safeguards. False positives can damage reputation, restrict expression, or disproportionately affect communities whose visual data are underrepresented in training sets. Platforms should conduct subgroup and cross-domain audits, monitor error rates after model updates, retain decision logs, provide meaningful appeals, and disclose when automated tools contributed to an action. Data minimization and retention limits are necessary because facial-video monitoring involves biometric and potentially sensitive information. These requirements are consistent with human-rights-centered principles of transparency, fairness, accountability, privacy, and human oversight in AI governance (Unesco, 2021).

Evidence from adjacent applied-AI domains also illustrates the broader value of combining deep representations with optimization and hybrid modeling strategies. Aliabadian integrated deep learning with ant colony optimization for melanoma classification, while Torabi et al. used particle swarm optimization in a deep-learning system for skin melanoma detection (Aliabadian, 2025; Torabi et al., 2024). Mohammadi and Anisheh applied deep learning to sensor-based human activity recognition (Mohammadi & Anisheh, 2024), and Karimi Dastgerdi and Zamani Boroujeni reviewed deep-learning and hybrid architectures for financial-market prediction (Karimi Dastgerdi & Zamani Boroujeni, 2020). Although these studies address different application domains and do not directly validate deepfake detection, they demonstrate the wider capacity of hybrid and optimization-guided AI to improve feature extraction, parameter search, and generalization in complex classification tasks. This cross-domain evidence supports the design rationale for combining complementary backbones, while the effectiveness of the present system must still be judged primarily through deepfake-specific and social-media-specific evaluation.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

5.3. *Limitations*

Several limitations should be acknowledged. First, the study was frame-based and did not model temporal or audio-visual consistency. Second, repeated-run variability, confidence intervals, and significance tests were not reported. Third, the available split description does not confirm strict video-level or identity-level separation. Fourth, the Visibility Matrix lacks a complete mathematical specification and controlled ablation. Fifth, exact dataset and partition counts, face-processing procedures, compression settings, and the unseen-data protocol were not fully documented. Sixth, several claims regarding reduced false positives and improved fake recognition were reported without raw counts or confusion matrices. Finally, platform-scale latency, video-level aggregation, calibration, subgroup performance, reviewer behavior, and downstream moderation effects were not evaluated.

Future work should incorporate temporal architectures, explicit video-level aggregation, audio-visual consistency, generator-disjoint evaluation, stronger cross-dataset testing, calibration analysis, and interpretable evidence maps. Compression-aware augmentation and visibility learning should be compared with restoration, frequency-domain, and self-consistency baselines under a shared protocol.

6. Conclusion

This study evaluated a frame-level deepfake detector that combines ResNet50 and EfficientNet-B0 features and incorporates iterative Visibility Matrix training. The fused model produced the highest reported point estimates, including 83% validation accuracy, AUC-ROC of 0.89, balanced accuracy of 0.81, and approximately 18 ms model inference per image. Preliminary results also suggested retained accuracy under an additional unseen evaluation and a severe-compression condition. These findings are not sufficient to establish compression resilience, cross-dataset generalization, or platform readiness because the split protocol, Visibility Matrix algorithm, compression procedure, and external evaluation were not fully specified and no repeated-run or ablation evidence was available. The system should therefore be treated as a candidate frame-level triage method requiring independent, video-disjoint validation and human oversight.

Authors' Contributions

Samir Qaisar Ajmi conducted the implementation, experiments, data preparation, visualization, and initial drafting. Mahmood Deypir contributed to conceptualization, methodology, supervision, validation, and critical revision. Razieh Farazkish contributed to validation, interpretation, and manuscript review. Ali Broumandnia contributed to supervision, methodological review, and critical revision. All authors reviewed and approved the final manuscript and accept responsibility for its content.

Declaration

Generative AI-assisted tools were used solely to support English-language editing, organization, and formatting of the manuscript. They were not used to create experimental data, perform model training, calculate the reported results, or alter the study findings. The authors reviewed and approved the complete manuscript and remain responsible for its accuracy and integrity.

Transparency Statement

FaceForensics++ and Celeb-DF v2 are publicly available research benchmarks subject to their respective access terms. Independent verification requires release of the exact video and frame lists, split indices, preprocessing and face-processing code, random seeds, Visibility Matrix implementation, compression scripts, hyperparameters, trained weights, and raw prediction outputs. These materials were not included with the manuscript.

Acknowledgments

The authors acknowledge the developers and maintainers of the FaceForensics++ and Celeb-DF datasets and the open-source deep-learning libraries used in the experimental workflow.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

The study involved secondary computational analysis of publicly available benchmark datasets and did not directly recruit, contact, or intervene with human participants. No new biometric recordings were collected by the authors. Dataset use should remain consistent with the licenses, access conditions, and research purposes established by the dataset providers. The proposed application to social media monitoring raises additional ethical obligations beyond dataset compliance. Deepfake scores should not be treated as definitive judgments about a person, speaker, or account, and the model should not be used for automatic censorship, punitive action, or covert population surveillance. Any operational implementation should apply data minimization, secure processing, limited retention, subgroup performance auditing, calibrated uncertainty, human review, transparent notification, and a meaningful appeal process. Particular care is required in political, journalistic, legal, and reputational contexts, where false-positive decisions may produce serious social harm. Deployment should follow human-rights-centered principles of privacy, fairness, accountability, transparency, proportionality, and human oversight (Unesco, 2021).

References

Aliabadian, A. (2025). Melanoma skin cancer detection using deep learning and the ant colony optimization algorithm. *Transactions on Machine Intelligence*, *8*(4), 181-190. https://doi.org/10.22034/tmi.2025.244866

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753-1820. https://doi.org/10.15779/Z38RV0D15J

Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2021). ID-Reveal: Identity-aware DeepFake video detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The Deepfake Detection Challenge dataset. arXiv:2006.07397,

Gao, J., Xia, Z., Marcialis, G. L., Dang, C., Dai, J., & Feng, X. (2024). DeepFake detection based on a high-frequency enhancement network for highly compressed content. *Expert Systems with Applications*, *249*, 123732. https://doi.org/10.1016/j.eswa.2024.123732

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. https://doi.org/10.12987/9780300235029

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Han, B., Han, X., Zhang, H., Li, J., & Cao, X. (2021). Fighting fake news: Two-stream network for deepfake detection via learnable SRM. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(3), 320-331. https://doi.org/10.1109/TBIOM.2021.3065735

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Karimi Dastgerdi, A., & Zamani Boroujeni, F. (2020). A review of deep learning methods for financial market prediction. *Transactions on Data Analysis in Social Science*, *2*(3), 164-172. https://doi.org/10.47176/TDASS.2020.164

Ke, J., & Wang, L. (2023). DF-UDetector: An effective method towards robust deepfake detection via feature restoration. *Neural Networks*, *160*, 216-226. https://doi.org/10.1016/j.neunet.2023.01.001

Kim, M., Tariq, S., & Woo, S. S. (2021). FReTAL: Generalizing deepfake detection using knowledge distillation and representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Le, B. M., & Woo, S. S. (2023). Quality-agnostic deepfake detection with intra-model collaborative learning. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Lee, S., An, J., & Woo, S. S. (2022). BZNet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos. Proceedings of the ACM Web Conference,

Li, Y., Bian, S., Wang, C., Polat, K., Alhudhaif, A., & Alenezi, F. (2023). Exposing low-quality deepfake videos of social network services using a spatial restored detection framework. *Expert Systems with Applications*, *231*, 120646. https://doi.org/10.1016/j.eswa.2023.120646

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Lin, H., Huang, W., Luo, W., & Lu, W. (2023). DeepFake detection with multi-scale convolution and vision transformer. *Digital Signal Processing*, *134*, 103895. https://doi.org/10.1016/j.dsp.2022.103895

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *Acm Computing Surveys*, *54*(1), 1-41. https://doi.org/10.1145/3425780

Mohammadi, S., & Anisheh, S. M. (2024). Human activity recognition based on deep learning using sensor data. *Transactions on Data Analysis in Social Science*, *6*(4), 222-230. https://doi.org/10.47176/TDASS.2024.222

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Perera, A. S., Atukorale, A. S., & Kumarasinghe, P. (2022). Employing super resolution to improve low-quality deepfake detection. Proceedings of the 22nd International Conference on Advances in ICT for Emerging Regions,

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision,

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision,

Sun, Z., Han, Y., Hua, Z., Ruan, N., & Jia, W. (2021). Improving the efficiency and robustness of deepfakes detection through precise geometric features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning,

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

Torabi, T., Esmaeili, K., Omran, M., & Anisheh, S. M. (2024). Skin melanoma cancer detection using particle swarm optimization algorithm and deep learning. *Transactions on Machine Intelligence*, *7*(3), 170-178. https://doi.org/10.47176/TMI.2024.170

Unesco. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Verdoliva, L. (2020). Media Forensics and Deepfakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910-932. https://doi.org/10.1109/JSTSP.2020.3002101

Wang, B., Wu, X., Tang, Y., Ma, Y., Shan, Z., & Wei, F. (2023). Frequency domain filtered residual network for deepfake detection. *Mathematics*, *11*(4), 816. https://doi.org/10.3390/math11040816

Woo, S. S. (2022). ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. Proceedings of the AAAI Conference on Artificial Intelligence,

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *IET Biometrics*, *10*(6), 607-624. https://doi.org/10.1049/bme2.12031

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433