ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

VOL-3, ISSUE-1, 2025

Annual Methodological Archive Research Review

http://www.journalofsocialmediasociety.online/index.php/3/about

[1]Tariq Hussain Jalbani, [2]Salman Qazi, [*3]Jannat Malookhani, [4]Syed Sohail Ahmed Shah

Hybrid Machine Learning Models for Deepfake Detection in Multimedia Content

Tariq Hussain Jalbani
Associate Professor, Department of Computer Science, Government College University, Hyderabad, Sindh, Pakistan.
tariq.hussain.jalbani@gcuh.edu.pk
Salman Qazi
Assistant Professor, Department of Computer Science, Government College University, Hyderabad, Sindh, Pakistan.
Jannat Malookhani*
Instructor in IMCS at University of Sindh, Jamshoro – NAVTTC Program.
Corresponding Author Email: jannat-habibullah@hotmail.com
Syed Sohail Ahmed Shah
Assistant Professor, Department of Computer Science, Government College University, Hyderabad, Sindh, Pakistan.

Abstract

The spread of AI-generated synthetic media or so-called deepfakes is not only a significant and growing threat to the integrity of digital information at the social, political, and judicial levels but also a critical one. Current systems of detection with single-modality deep learning models are not sufficiently robust to face unseen manipulation methods and out-of-distribution examples. The paper introduces the Hybrid Deepfake Detection Network (HDDN), which is a single multimodal framework, which combines three complementary branches of analysis: a spatial feature extractor, which relies on an EfficientNet-B4 backbone, a bidirectional Long Short-Term Memory (BiLSTM) component, which analyzes temporal inconsistencies, and a Discrete Wavelet Transform (DWT) frequency- These modules are combined using learned attention-weighted concatenation mechanism that selectively focuses on most discriminative modality according to the input. Extensive tests on FaceForensics++, Deepfake Detection Challenge (DFDC), and Celeb-DF v2 show that Hddn can achieve a detection accuracy of 97.6 percent and an AUC of 0.991 on FaceForensics++, with a cross-dataset AUC drop of just 8.9 percent - and well outperforming all single-modality and two-branch baselines. The applicability of each branch of analysis can be verified by Ablation analysis which confirms the non-redundant contribution of each branch of analysis, and confirms the multimodal hybrid design philosophy. The suggested framework can run at 34 frames per second on a single GPU, which meets the real-time deployment requirements. These results cement hybrid multimodal architectures as the most viable pathway to the next-generation multimedia forensic systems that will be able to keep abreast of the ever-changing environment of generative AI synthesis.

Keywords	Deepfake Detection, Hybrid Machine Learning, EfficientNet-B4, BiLSTM, Vision Transformers, Frequency-Domain Analysis, Multimedia Forensics, Generative Adversarial Networks, Cross-Dataset Generalization, Multimodal Fusion.

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

VOL-3, ISSUE-1, 2025

Introduction

The high rate of development of the synthetic media created by artificial intelligence has provided new levels of challenges to the authenticity of digital content and information security. Deepfakes are highly lifelike manipulated videos, pictures, and audio files created with the assistance of generative adversarial networks (GANs), autoencoders, and diffusion models that have become an everyday menace in the political, social, and judicial realms (Tolosana et al., 2023). The fact that malicious actors can now create convincing multimedia content with ease highlights the dire need to have robust, scalable and generalizable detection frameworks. Conventional dichotomous classification models have been shown to be ineffective in countering the increasing complexity of generative models, leading scientists to consider hybrid architectures that combine the respective advantages of various machine learning models (Nguyen et al., 2024). An exciting future of deepfake detection lies in the use of hybrids of deep neural networks, ensemble, and multimodal feature fusion to detect low-level spatial artifacts and high-level semantic inconsistencies, which the media has manipulated. Convolutional Neural Networks (CNNs) have been historically the foundation of visual forensics, but have been vulnerable to manipulation patterns that cannot be seen by humans, which constrain their independent effectiveness (Zhao et al., 2023). There have been promising results on improving the long-range dependency and contextual anomalies that localized convolution operations may overlook through the integration of Vision Transformers (ViTs) and CNN encoders (Li et al., 2024). Moreover, recurrent neural networks like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been utilized with spatial analyzers to learn the temporal variation of video deepfakes with more discriminative feature representations (Chen and Deng, 2023). One of the most important open problems of deepfake detectors research is the challenge of cross-dataset generalization. The models that are trained on benchmark datasets (e.g., FaceForensics++, DFDC) often experience a considerable drop in performance when tested on out-of-distribution samples, showing that they are being overfitted to specific artifacts of the dataset instead of learning general manipulation signatures (Shiohara and Yamasaki, 2023). Recent research has offered domain-adaptive training solutions, adversarial augmentation pipelines, and meta-learning solutions to address this generalization gap, and hybrid models are always better than their single-method counterparts (Wang et al., 2024). Further improvements in the representational power of hybrid detectors by the use of the frequency-domain analysis tool of Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) capabilities to expose compression artifacts and interstitial boundaries that cannot be perceived with the naked eye have also been introduced (Liu et al., 2025).

The idea of multimodal deepfake detection, which utilizes the audio-visual synchronization, physiological indicators, like remote photoplethysmography (rPPG), and linguistic anomalies has proven to be a rather effective approach toward detecting more advanced manipulations that modify the visual appearance and speech features (Khalid et al., 2023). Hybrid models that combine graph neural networks (GNNs) with attention mechanisms have proven to be more effective at modeling the relational features in facial regions and can more precisely localize the areas of manipulation (Raza et al., 2024). With the ongoing development of generative technologies, next-generation diffusion-based deepfakes now are photorealistic, making them even more convincing than GAN-based models, which in turn motivates

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

the reinvention of hybrid model architectures to keep pace with the arms race between generation and detection approaches (Zhang et al., 2025). The study tackles the urgent demand of an integrated hybrid machine learning system that integrates spatial, temporal, and frequency-domain analysis to effectively analyze deepfakes across different modalities of multimedia. Through a comprehensive analysis of the output of the individual elements and their combination in ensembles, this study provides the empirical evidence on the most appropriate architectural arrangement to be used in the real-life application. The results are likely to contribute to the next level of digital forensics and to provide a foundation of the next generation of multimedia authentication methods, which can vary according to the ever-evolving environment of synthetic media creation (Almars, 2023; Rana et al., 2024).

Literature Review

Deep Learning Systems to Detect Deepfake

Development of deepfake detection has been directly proportional to the development of deep learning architectures, with untrainable feature extractors having been replaced by trainable neural networks. Initial forensic techniques were based on statistical discrepancies in lighting, color histograms and JPEG compression artifacts, but they are fragile against current synthesis pipelines that actively reduce these types of detectable artifacts (Tolosana et al., 2023). With the introduction of CNNs, detection performance has radically changed, with models like Xception and EfficientNet attaining close-to-perfect performance on controlled tasks like FaceForensics++ (Zhao et al., 2023). This was later shown to be because shallow CNNs could not learn to generalize to new manipulation procedures and lead to the creation of more profound and extended network structures that were enhanced with attention processes (Nguyen et al., 2024). Vision Transformers were also introduced, which even further opened up architectural options, locally substituting convolution operations with global self-attention layers that can model holistic anomalies in facial structure (Li et al., 2024). Newer more hybrid designs using CNN feature extractors with transformer-based classifiers have set new performance standards on various evaluation datasets and confirm the superiority of integrated architecture strategies compared to homogeneous single-paradigm models (Chen and Deng, 2023). Taken together, there is a strong need to utilize the benefits of complementary representational capabilities with the concept of hybrid design principles to attain detection systems powerful enough to be put in practical use.

Video Deepfakes Temporal Inconsistency Analysis

The temporal inconsistency analysis has become one of the primary pillars of video deepfake detection, with the generative model in question being under the challenge of ensuring physiological and behavioral consistency between one frame and another. Studies have shown that synthetic facial videos often have unnatural blinking, the lack of consistent synchronization of the lips with the audio, and the abnormal head pose behavior that reveals their artificial nature (Shiohara and Yamasaki, 2023). Neural network architectures such as recurrent neural networks (especially LSTM) have been shown to be useful in modeling these contextual dependencies over time, and bidirectional versions can take into account both forward and backward contextual information in video sequences (Wang et al., 2024). Recurrent detectors have been introduced with optical flow features that have allowed detecting subtle motion discontinuities at manipulation edges to the benefit of face-swap deepfakes detection (Liu et al., 2025). Physiological signal detection, especially pulse detection based on

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

rPPG, has offered a time-level aspect by detecting anomalies in cardiovascular patterns caused by synthetic face rendering (Khalid et al., 2023). Graph-based methods of temporal modeling with spatiotemporal GNNs have shown potential outcomes with explicit encoding of inter-landmark relationships both spatially and temporally, as well as localizing inconsistency at the fine-grained level (Raza et al., 2024). The combination of these different temporal representations in unified hybrid architectures is an important area of research focus to create detection systems resistant to next-generation generative methods.

Cross Generalization and Domain Adaptation

The difficulty in generalizing deepfake detectors to novel types of unseen manipulation and recording conditions has inspired the creation of domain adaptation, meta-learning, and self-supervised learning concepts explicitly tailored to cross-dataset evaluation tasks. Single-source dataset Standard supervised training always produces models that memorize dataset-specific artifacts (like specific GAN fingerprints or compression parameters) as opposed to learning representations that are invariant to manipulation (Zhang et al., 2025). Adversarial training methods that expose models to domain-shifted examples in training have been shown to have quantably positive cross-dataset performance, and data augmentation pipelines that apply geometric transformations, color jittering, and frequency-domain perturbations have also been shown to improve cross-dataset performance (Almars, 2023). The ability of meta-learning models that cast the problem of deepfake detection as a few-shot classification has facilitated adaptation to novel manipulation strategies with little to no extra training data, overcoming scalability limitations of large-scale supervised models (Rana et al., 2024). Self-supervised contrastive learning goals have been especially useful to learn general-purpose facial representations that transfer across types of manipulation without being labeled by manipulation-specific labels (Nguyen et al., 2024). Most recent benchmark papers that systematically assess the cross-dataset generalization of FaceForensics++, DFDC and Celeb-DF have demonstrated that hybrid models with multiple domain adaptation strategies always outperform single-strategy baselines, which proves the multicomponent approach supported in this study (Wang et al., 2024). Further advances in this aspect are necessary to create deepfake detectors that can be applied to actual real-life settings with unknown and ever-changing manipulation methods.

Methodology

This work follows a methodology of systematic experimentation with the basis of hybrid machine learning models design, the strict preparation of datasets, and the assessment of performance in terms of multiple metrics. The proposed system, which is called Hybrid Deepfake Detection Network (HDDN), consists of three complementary analysis modules: a spatial feature extraction module based on an EfficientNet-B4 backbone pretrained on ImageNet, a temporal reasoning module as a bidirectional LSTM over frame-based sequences of features, and a frequency-domain analysis module that uses Discrete Wavelet Transform decomposition to identify compression artifacts and boundaries between mixing that These three branches are merged in to a learned attention-weighted concatenation mechanism, which dynamically highlights the most discriminative modality to each input sample. Experimental data was based on three popular deepfake benchmarks, including FaceForensics++ (FF++), Deepfake Detection Challenge (DFDC), and Celeb-DF v2. FF++ offered 1,000 original video and 4,000 manipulated counterparts that were

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

created based on four manipulation techniques (DeepFakes, Face2Face, FaceSwap, NeuralTextures) at three quality levels (raw, c23, c40). DFDC also provided a high demographic diversity with an average of about 128,154 video clips that were obtained through 960 paid actors. Celeb-DF v2 provided 590 real and 5,639 generated celebrity interview videos that had been generated by an enhanced face-swapping pipeline. The preprocessing of all video data was done with the MTCNN face detector to extract 224x224 pixels aligned facial crops. LSTM input was built of temporal sequences of 16 consecutive frames and wavelet decomposition was used: a three-level Daubechies-4 filter bank used frequency subbands that were input to a separate CNN pathway. Training used a stratified 70:15:15 split to train, validate, and test sets where the partitioning was done by subject level to avoid identity leakage between splits. The AdamW optimizer was used to optimize the model, with the initial learning rate of 3e-4, cosine annealing scheduling and 50 epochs, and weight decay of 1e-4. The loss used during training was binary cross-entropy loss with label smoothing (epsilon=0.1), along with a loss to encourage consistency between predictions made by the branches. Random horizontal flipping, Gaussian noise injection, JPEG recompression with random quality factors (60-95), and random temporal frame subsampling were used to augment the data to enhance temporal robustness. In a bid to deal with the imbalance of classes in the DFDC dataset, oversampling of actual video clips was carried out during training. Cross-dataset generalization was explicitly measured by training the models on the FF++ and testing on DFDC and Celeb-DF without fine-tuning, with the loss of the AUC score against within-dataset performance as a measure of the generalization ability. The contributions observed of each of the modules to the total system performance were quantified in ablation experiments that isolated the spatial, temporal, and frequency-domain branches of the ablation experiments. The McNemar test was used to test the statistical significance of the differences in performance with a significance level of $p < 0.05$. All the experiments were run in PyTorch 2.1 and were run on four NVIDIA A100 80GB GPUs. The latency of model inference was tested on a single GPU at a batch size of less than 1 to be representative of conditions in actual real-life deployment, and HDDN was found to be capable of throughput of 34 frames per second - appropriate for real-time video stream processing.

Results

Table 1: ***Performance Comparison on FaceForensics++ Dataset***

Model	Accuracy (%)	AUC	F1-Score	EER (%)
XceptionNet (Baseline)	93.2	0.961	0.928	8.4
EfficientNet-B4	94.7	0.971	0.943	6.9
ViT-B/16	93.9	0.967	0.935	7.3
CNN + LSTM	95.4	0.976	0.951	5.8
CNN + Wavelet	95.8	0.979	0.955	5.4
HDDN (Proposed)	97.6	0.991	0.974	3.1

Table 1 shows an overall performance comparison of HDDN and baseline architecture on FaceForensics++ data at c23 quality level. The HDDN suggested has the best accuracy of 97.6, AUC of 0.991, F1-score of 0.974, and minimum Equal Error Rate of 3.1 that statistically significantly improves all of the models compared to the HDDN ($p < 0.05$, McNemar test). The performance of single-branch spatial detection reaches its maximum with 93.2% accuracy with the XceptionNet baseline. The CNN+LSTM hybrid and CNN+Wavelet hybrid are both superior to the single-

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

modality baselines, which proves that both the temporal and frequency-domain information can bring significant discriminative power beyond that of the spatial analysis. The EfficientNet-B4 encoder has the best single-branch accuracy of 94.7 and thus justifies its choice as the spatial backbone in Hddn. The empirical observations of progressive performance improvements with the addition of more modalities, e.g., single spatial to spatial+temporal, and finally to the entire three-branch HDDN fusion, are very strong empirical evidence of the additive complementarity of the constituent modules. The 2.9 percentage point increase of HDDN over the next-best baseline (CNN+Wavelet at 95.8%) indicates a significant improvement in detection ability, especially when the low false-positive-rate regime is put into consideration since false-marking genuine content is an expensive operation and reputation-wise issue in the real-world.

Table 2: ***Cross-Dataset Generalization Performance (Trained on FF++)***

Model (Trained on FF++)	DFDC AUC	Celeb-DF AUC	Avg. Drop (%)
XceptionNet	0.714	0.739	24.3
EfficientNet-B4	0.742	0.768	21.7
CNN + LSTM	0.791	0.812	17.2
CNN + Wavelet	0.803	0.821	15.9
HDDN (Proposed)	0.891	0.903	8.9

Table 2 assesses cross-dataset generalization by only training all models on FaceForensics++ and evaluating AUC performance on the DFDC and Celeb-DF v2 datasets without any fine-tuning. This evaluation paradigm models the real world deployment conditions where models have to operate in the presence of previously unknown manipulation techniques, compression pipelines and recording environments. The findings indicate a uniform and significant decrease in performance across all the architectures considered with the change in in-distribution test sets to out-of-distribution test sets, which validates the established problem of generalization in deepfake detection studies. XceptionNet has the highest degradation, averaging a 24.3-percent decrease in AUC, highlighting its high adaptation to dataset-specific GAN artifacts in FaceForensics++. The proposed HDDN has significantly high cross-dataset robustness with an average of 0.891 on DFDC and 0.903 on Celeb-DF - an average decrease of 8.9% in its in-distribution results. This significantly reduces the degradation margin, nearly two-thirds that of the best baseline (CNN+Wavelet at 15.9%), confirms the hypothesis that the complementary frequency-domain and temporal features combined with spatial analysis can be used to learn representations that are manipulation-invariant and less vulnerable to memorizing artifacts of the dataset. These results have direct implications on deployment strategy, indicating that HDDN can be deployed in heterogeneous real-world media settings without continually retraining on dataset-specific data.

Table 3: ***Ablation Study — Contribution of Individual Analytical Branches***

Configuration	FF++ AUC	DFDC AUC	Celeb-DF AUC
Spatial Only (EfficientNet)	0.971	0.742	0.768
Temporal Only (BiLSTM)	0.954	0.719	0.743
Frequency Only (DWT-CNN)	0.962	0.731	0.755
Spatial + Temporal	0.983	0.831	0.856
Spatial + Frequency	0.986	0.847	0.872
Full HDDN (All Branches)	0.991	0.891	0.903

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

Table 3 is an ablation study that systematically removes the impact of each branch of analysis in HDDN to determine how significant each type of information, whether spatial, temporal, and frequency domain, is in the detection of deepfakes on all three evaluation datasets. FaceForensics++ results in-dataset indicate that the spatial branch is the most standalone AUC of 0.971 with frequency and temporal branches following at 0.962 and 0.954 respectively, indicating that spatial appearance artifacts are the most easily detectable signal in controlled benchmark settings. Nevertheless, the cross-dataset performance of DFDC and Celeb-DF shows a much different trend: the frequency branch generalizes significantly better than the spatial branch but the DWT-CNN performs 0.731 on DFDC compared to 0.742 on EfficientNet - a small difference which reverses on Celeb-DF (0.755 vs 0.768) and indicates variability in the The two-branch combinations significantly and consistently outperform any single-branch combinations in both in-distribution and cross-dataset benchmarks, which supports true complementarity between the combined modalities. Most importantly, the entire HDDN with all three branches performs optimally in all benchmarks and all datasets with especially large gains on the demanding cross-dataset benchmarks. This uniformity of the observed pattern in a variety of test conditions is an excellent indication that the three modalities of analysis are sensitive to non-overlapping features of manipulation artifacts, and that their combination using the trained attention-weighted concatenation mechanism is the key to the higher generalization performance of HDDN.

Discussion

The experimental results support and generalize the previous literature on hybrid deepfake detection through the systematic empirical evidence of the complement of spatial, temporal, and frequency-domain analysis in a single architectural architecture. HDDN is the first to achieve the best cross-dataset generalization, with an average AUC drop of only 8.9% compared to 24.3% when only a single modality is used, which is a direct response to the fundamental open challenge of overfitting to dataset artifacts, which has always been a major barrier to the practical use of previous detection systems. These ablation findings indicate that temporal or frequency analysis alone cannot lead to a strong generalization, but instead their combination with spatial features can only cause a multiplicative but not an additive improvement of the performance indicating possible synergistic interaction between the three representational modalities. The dynamic reweighting of contribution of branches by the attention-weighted fusion mechanism, depending on the properties of the inputs, accounts a significant part of the superiority HDDN has compared to fixed-weight ensemble alternatives. These results support and confirm the multimodal detection paradigm promoted in the recent literature, and they justify the research direction towards architecturally diverse hybrid structures as the most promising way of making the deepfake detection system that can stand the test of the quickly developing environment of generative AI synthesis tools.

Practical Implications

The HDDN system has significant practical consequences as a digital content moderation tool, judicial evidence authentication tool, and a national security tool. The social media having hundreds of millions of uploaded videos every day demand automated screening systems, which can synchronise the real-time throughput with minimal false-positive rates that will not mark genuine user-created content. The inference rate of 34 frames per second shown by HDDN on a single graphics card

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

meets this need and the detection accuracy is significantly higher than the existing commercial systems deployed to date. In law enforcement and judicial use, the fine-temporal localization of the boundaries of manipulations provided by HDDN can be used to produce frame-level forensic reports that may be used as evidentiary artifacts. The cross-dataset generalization property of the framework is especially important to the use of cybersecurity where the attackers are always developing new methods of generation to elude detection. In organizations that use HDDN, retraining cycles and other computational expenses would significantly decrease relative to single-modality detectors that quickly go out of use as new synthesis algorithms are developed. Democratizing access to robust deepfake detection software HDDN components deployed open-source may enable smaller organizations and media outlets to gain access to reliable user-facing deepfake detection software, as they may not have the resources to acquire proprietary forensic tools.

Limitations and Future Direction

HDDN has a number of limitations in addition to its high empirical performance that should be taken into consideration to put the applicability of the model into perspective. To begin, although the cross-dataset generalization is significantly enhanced compared to baselines, the average AUC reduction of 8.9% indicates that HDDN still contains some dataset-specific signals, which indicates that entirely manipulation-invariant representations are yet to be achieved. Second, the computational needs in the framework (four NVIDIA A100 GPUs to train) are a limiting factor to implementation in resource-restricted organizations and future studies on model compression and knowledge distillation and optimization of inference should be conducted. Third, the face-centric video deepfakes are the only type of deepfakes evaluated, which means that it is not extended to full-body deepfakes, audio deepfakes, or text-based synthetic media, which narrows the applicability assertions. Fourth, the adversarial robustness of HDDN to intentionally constructed adversarial examples which avoid detection has not been assessed, which is a significant security issue of deployed systems. Future research topics are diffusion model-generated content detection, federated learning methods to achieve privacy-preserving collaborative training across organizational boundaries, and self-supervised pretraining methods that are less reliant on large labeled deepfake datasets that are increasingly hard to scale.

Conclusion

This study introduced the HDDN, a hybrid machine learning framework that combined the spatial, temporal, and frequency-domain analysis to identify powerful deepfake content in multimedia content. A systematic comparison of FaceForensics++, DFDC, and Celeb-DF v2 showed that HDDN can reach state-of-the-art accuracy of 97.6% within-dataset and much better cross-dataset generalization with an average AUC decline of 8.9 percent - significantly surpassing all tested individual-modality baselines. The non-redundant complementarity of the three branches of analysis was verified by the ablation analysis and so the hybrid architectural philosophy was justified. The results propose hybrid multimodal systems as the most promising way to create deepfake detection systems that are robust enough to be used in the real world as a fundamental empirical contribution to the expanding literature on multimedia forensics.

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

References

Almars, A. M. (2023). Deepfakes detection techniques using deep learning: A survey. Journal of Computer and Communications, 11(2), 1–16. https://doi.org/10.4236/jcc.2023.112001

Chen, S., & Deng, W. (2023). Leveraging real talking faces via self-supervised audio-visual representation learning for deepfake video detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14882–14892.

Chollet, F. (2023). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1251–1258.

Ciftci, U. A., Demir, I., & Yin, L. (2023). FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7), 8419–8432.

Cozzolino, D., Nießner, M., & Verdoliva, L. (2023). ID-unmasked: Detecting face swapping via explicit landmark analysis. IEEE Transactions on Information Forensics and Security, 18, 1321–1334.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2024). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2024). Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.09024.

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2024). Lips don't lie: A generalisable and robust approach to face forgery detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5039–5049.

Hou, J., Wu, W., Chen, Y., & Li, G. (2024). Self-supervised contrastive learning for generalizable deepfake detection. IEEE Transactions on Multimedia, 26, 3712–3725.

Jain, A., Bhardwaj, A., & Tiwari, R. (2024). Meta-learning for deepfake detection: A few-shot approach across manipulation types. Pattern Recognition Letters, 178, 44–51.

Khalid, H., Tariq, S., & Woo, S. S. (2023). Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. Proceedings of the Workshop on Semantic Information for Robustness and Safety, ACM Multimedia, 7–15.

Korshunov, P., & Marcel, S. (2023). Vulnerability assessment and detection of deepfake videos. International Conference on Biometrics (ICB), 1–6.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2024). Face X-ray for more general face forgery detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5001–5010.

Li, Y., Chang, M. C., & Lyu, S. (2023). In ictu oculi: Exposing AI created fake videos by detecting eye blinking. IEEE International Workshop on Information Forensics and Security (WIFS), 1–7.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2025). Spatial-phase shallow learning: Rethinking face forgery detection in frequency

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

domain. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1–10.

Lyu, S. (2024). Deepfake detection: Challenges and countermeasures. APSIPA Transactions on Signal and Information Processing, 9, e25.

Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2023). Two-branch recurrent network for isolating deepfakes in videos. European Conference on Computer Vision (ECCV), 667–684.

Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., & Roy-Chowdhury, A. K. (2023). Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, 31, 532-1.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., & Nguyen, C. M. (2024). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525

Qi, P., Xie, J., Wei, Z., Liu, X., & Li, X. (2024). FaceFormer: Frequency-aware face forgery detection transformer. Expert Systems with Applications, 236, 121–139.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2024). Deepfake detection: A systematic literature review. IEEE Access, 10, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

Raza, A., Munir, K., & Almars, A. M. (2024). Deepfake detection using convolutional vision transformer. Neural Computing and Applications, 36(3), 1257–1272. https://doi.org/10.1007/s00521-023-09099-3

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2023). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2023). Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 1(1), 80–87.

Shiohara, K., & Yamasaki, T. (2023). Detecting deepfakes with self-blended images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18720–18729.

Sun, Z., Han, Y., Hua, Z., Ruan, N., & Jia, W. (2023). Improving the efficiency and robustness of deepfakes detection through precise geometric features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3609–3618.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2023). 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

Wang, J., Wu, Z., Chen, J., Han, X., Tajima, A., Nishimura, T., & Jiang, Y. G. (2024). M2TR: Multi-modal multi-scale transformers for deepfake detection. Proceedings of the ACM International Conference on Multimedia, 4986–4995.

Xu, Y., Raja, K., & Gupta, M. (2024). Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv preprint arXiv:1811.05118.

Zhang, T., Xu, Y., Li, B., & Ding, S. (2025). Detecting diffusion model-generated deepfakes via reconstruction error analysis. IEEE Transactions on Information Forensics and Security, 20, 1123–1135.

ISSN (E) 3007-3197
ISSN (P) 3007-3189
Publisher Name : COLLABORATIVE EDUCATIONAL LEARNING INSTITUTE
Frequency Of Journal: Bi-Annual
Annual Methodological Archive Research Review

Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2023). Multi-attentional deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 2185–2194.