A Hybrid Approach for Robust Deep Fake Image Detection using Spatial and Frequency Domain Features

Uma Yadav
School of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Ramdeobaba University, Nagpur, India
uma.yadav12@gmail.com (corresponding author)

Priya Dasarwar
Department of Computer Science and Engineering, Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India
priyadasarwar21@gmail.com

Shweta Bondre
School of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Ramdeobaba University, Nagpur, India
shwetakharat1510@gmail.com

Supriya Kalamkar
Electronics & Telecommunication, Army Institute of Technology, Pune, India
shadke.9@gmail.com

Received: 4 February 2025 | Revised: 24 February 2025, 15 March 2025, and 19 March 2025 | Accepted: 22 March 2025

 DOI: https://doi.org/10.48084/etasr.10458

ABSTRACT

The rise of deepfake technology has transformed media synthesis, enabling the creation of hyperrealistic yet manipulated images and videos. Although these innovations offer creative opportunities, they also introduce severe risks such as misinformation, identity theft, and decreased trust in digital content. This study presents a hybrid approach to deepfake image detection that integrates features from the spatial and frequency domains to improve detection accuracy. The proposed method combines multiscale Convolutional Neural Networks (CNNs), frequency domain analysis, attention-based transformer networks, and ensemble learning to identify manipulation artifacts and enhance classification robustness. The model was tested on a large-scale dataset of 140,000 images, evenly divided between real and fake images, achieving a training accuracy of 93% and a testing accuracy of 88%. Using adversarial training and advanced feature extraction techniques, the proposed approach effectively detects subtle artifacts introduced during manipulation. The experimental results illustrate the model's ability to generalize across diverse manipulation methods and datasets, making it a scalable and reliable tool for real-world applications. This research emphasizes the significance of hybrid detection frameworks in addressing the complexities of synthetic media forensics and underscores the necessity of multidomain feature integration to address the growing challenges posed by deepfakes.

Keywords-deep fake detection; hybrid approach; spatial domain analysis; frequency domain analysis; Convolutional Neural Networks (CNNs); attention-based transformers; media forensics; synthetic media

I. INTRODUCTION

Recent years have seen significant advances in deep fake technology, a product of generative Artificial Intelligence (AI), which makes it possible to create artificially realistic images, movies, and audio [1]. Derived from the words "deep learning" and "fake," the phrase "deep fake" emphasizes its use of sophisticated neural networks to create data that remarkably accurately replicate real-world data. Deepfake technology, which was first made popular in creative and entertainment applications, has now permeated fields such as marketing, education, and politics, revolutionizing the creation and consumption of digital media. Algorithms such as diffusion models, variational autoencoders, and Generative Adversarial Networks (GANs) are at the heart of deepfake creation [2]. These algorithms create data that are frequently indistinguishable from real data utilizing enormous datasets and processing capacity. For example, GANs use a generator-discriminator system in which the discriminator assesses the authenticity of the synthetic samples produced by the generator, hence increasing the output quality iteratively [3]. In contrast, diffusion models provide more control and accuracy in creation by refining images by reversing a noise diffusion process.

Deepfakes have a lot of potential for useful applications. They are employed in the entertainment sector to improve visual effects, de-age actors, or produce lifelike computer-generated characters [4]. They also provide individualized training and virtual simulations in classrooms. However, there are also serious hazards associated with this technology. Concerns around identity theft, misinformation, privacy infringement, and the decline of public trust are raised by the ability to produce incredibly realistic yet fraudulent content. Examples of controlled media disseminating misleading narratives and political deepfakes highlight how urgent it is to address these issues [5]. Existing methods for deepfake detection rely heavily on spatial domain analysis or temporal inconsistencies in videos. However, these approaches often struggle with generalized performance across diverse datasets and manipulation techniques. Furthermore, as deepfake generation methods evolve, the artifacts traditionally used for detection become increasingly subtle, challenging the robustness of detection models.

Recent advances in deepfake detection have explored various methods, ranging from metadata analysis to deep learning-based visual forensics. In [6], a hybrid approach was proposed, which integrated metadata inconsistencies (e.g., EXIF data mismatches) with visual anomalies such as unnatural facial features, significantly improving detection accuracy. FaceForensics++ [7] provided a benchmark dataset for manipulated facial images, helping deep learning models distinguish real from altered content. In [8], frequency domain analysis was used to detect imperceptible manipulation artifacts, demonstrating superior performance over spatial-domain methods. In [9], a two-stream neural network approach was introduced to analyze both facial and non-facial features, enhancing detection by capturing local and global inconsistencies. Capsule-forensics [10] utilized capsule networks to improve robustness against pose variations, excelling at detecting 3D face manipulations. In [11], deferred neural rendering was introduced, highlighting the challenges posed by advanced generative techniques that improve image realism. The survey in [12] comprehensively analyzed classical and machine learning-based detection techniques, highlighting the growing sophistication of deepfakes and the need for adaptive detection strategies. These studies collectively highlight the evolution of deepfake detection, underscoring the importance of integrating diverse methods to develop robust and accurate detection systems.

These studies also collectively illustrate the evolution of deepfake detection, from leveraging neural networks to innovative approaches such as capsule networks and frequency domain analysis. Table I provides an overview of various recent deepfake detection techniques, highlighting their key contributions. These studies collectively showcase the progression of deepfake detection methods, from dataset development to advanced neural architectures, highlighting the need for continuous innovation against deepfake threats.

This study presents a hybrid approach to deepfake detection that leverages features from both the spatial and the frequency domains. By integrating multiscale feature extraction using CNNs, frequency domain analysis, and an attention-based transformer network, the proposed system aims to enhance the detection of manipulation artifacts. The approach is further reinforced through adversarial training, ensemble learning, and cross-domain transfer learning to improve robustness and generalization. The primary contributions of this work include:

- A comprehensive method that combines features from the spatial and frequency domains for improved detection accuracy.

TABLE I. OVERVIEW OF DEEPFAKE DETECTION TECHNIQUES

Year	Study	Approach	Key features	Contributions
2023	[13]	Patch-based deepfake detection	Utilizes entire face and face patches to detect manipulations, enhancing accuracy in occluded images.	Improved detection accuracy, especially in images with obstructions.
2023	[14]	Comparative analysis of deepfake algorithms	Reviews various deepfake creation and detection techniques, emphasizing the need for robust detection methods.	Comprehensive overview highlighting challenges and future research directions.
2023	[15]	Leverages deep learning for deepfake detection	Explores different methodologies to achieve cost-effective models with higher accuracy across various datasets.	Addresses generalizability and proposes solutions for improved detection accuracy.
2021	[16]	Hybrid approach combining metadata analysis and visual forensics	Detects inconsistencies in metadata (e.g., EXIF data, timestamps) and visual anomalies (facial artifacts, lighting mismatches).	Enhanced robustness by integrating multiple modalities for detection.
2020	[8]	Frequency domain analysis	Detects inconsistencies in the frequency spectrum using Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT)	More reliable detection of high-quality fakes that lack pixel-level artifacts
2019	[7]	FaceForensics++ dataset and deep learning framework	Large-scale benchmark dataset with real and manipulated facial images, deep learning model utilizing facial and non-facial features.	Improved training for robust detection systems, benchmark dataset widely adopted in deep fake research.

- An attention-based mechanism to refine feature maps and prioritize manipulation-prone regions.
- Experimental evaluation demonstrating high detection accuracy on a large-scale dataset of 140,000 images, comprising diverse real and fake samples.

II. DATASET

The dataset utilized in this study consists of 140,000 images, evenly divided into 70,000 real and 70,000 fake images, sourced from Kaggle [17] and used for image forgery detection [18]. The real images are authentic representations of individuals, captured under diverse conditions, including variations in lighting, backgrounds, and facial expressions, ensuring a rich diversity of visual features. In contrast, the fake images are generated using advanced techniques such as StyleGAN and DeepFake, encompassing various types of manipulations such as facial reenactment, attribute editing, and synthetic face generation. These manipulated images exhibit common artifacts, including blending inconsistencies, unnatural textures, and pixel-level discrepancies, making them ideal for training and evaluating detection models. All images were preprocessed to a uniform resolution of 256×256 pixels, with annotations categorizing them as either real or fake. This comprehensive dataset is designed to reflect real-world scenarios, enhancing the robustness and generalizability of the proposed hybrid deepfake detection model. Figure 1 shows sample images of real and fake persons.

Fig. 1. Dataset samples: (a) Real images, (b) Fake images.

III. PROPOSED METHOD

The method was designed to achieve high accuracy in detecting deepfake images by employing an image-only approach. The system integrates state-of-the-art deep learning techniques, leveraging spatial, frequency, and attention-based feature extraction for robust detection [19]. Figure 2 shows the proposed method, which consists of the following stages.

A. *Preprocessing*

The initial stage involves input collection and preprocessing, which is essential to ensure consistency and improve the quality of data used in the detection pipeline. The dataset images are sourced from diverse datasets to include a variety of input types and manipulation techniques. This diversity ensures robustness by exposing the model to a broad spectrum of real and manipulated images, including those with subtle or complex alterations. To maintain uniformity in pixel intensity values, all images undergo normalization. This process adjusts the intensity distribution, mitigating inconsistencies across datasets and improving model performance [20]. Images are resized to a fixed resolution, 256×256, to standardize the input dimensions. This resizing not only facilitates efficient processing by the model but also ensures compatibility across different stages of the pipeline. Filters are applied to eliminate noise or artifacts that could hinder feature extraction. At the same time, care is taken to preserve any manipulation artifacts embedded in the images, as these are critical for accurate deepfake detection. This comprehensive preprocessing stage ensures that the input data is optimized for subsequent feature extraction and analysis, laying a strong foundation for accurate and reliable detection.

B. *Multi-Scale Feature Extraction Using CNNs*

A multiscale CNN architecture was designed to extract features at different levels of abstraction. The CNN consists of five convolutional layers, each followed by batch normalization, ReLU activation, and max pooling. Lower layers capture fine-grained details such as textures and pixel-level inconsistencies, often indicative of manipulations. The higher layers detect global structural information, such as facial geometry and object boundaries, providing a comprehensive representation of the image. The CNN utilizes 3×3 filters in early layers for edge and texture detection, progressively increasing to 5×5 and 7×7 filters in deeper layers for broader feature abstraction. The final feature maps are flattened and passed through fully connected layers, followed by a softmax activation function for classification. The extracted spatial features serve as the foundation for subsequent fusion with frequency domain features.

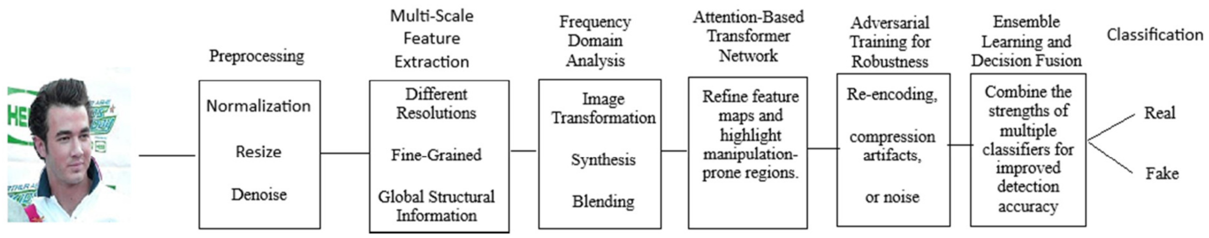


Fig. 2. Proposed architecture.

C. Frequency Domain Analysis

Frequency domain analysis is incorporated to detect imperceptible artifacts in the spatial domain. The system applies Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) to analyze high-frequency components, which often contain indicative signs of deepfake manipulations, such as stitching artifacts or unnatural blending. High-frequency residuals help detect minute inconsistencies introduced by generative models. The extracted frequency-based features are combined with spatial features from the CNN to provide a holistic understanding of the image's attributes. This fusion strengthens the model's ability to detect deepfakes, even in scenarios where spatial cues alone are insufficient.

D. Attention-Based Transformer Network

The attention-based transformer network refines the extracted features by dynamically focusing on manipulation-prone regions of the image [21]. By processing the concatenated spatial and frequency features, the network employs self-attention mechanisms to identify critical areas, such as edges, boundaries, or regions, with unnatural textures. This attention mechanism prioritizes relevant details while reducing redundant information, resulting in an improved feature map. The transformer network consists of six attention layers, each using multihead attention to capture dependencies across different image regions. Layer normalization and residual connections enhance feature stability, leading to a refined feature map that improves the model's ability to highlight subtle manipulation traces.

E. Adversarial Training for Robustness

To enhance the model's resilience against countermeasures, adversarial training is employed as a critical component of the method. This involves augmenting the training dataset with manipulated images that include various post-processing techniques, such as re-encoding, compression artifacts, or noise. When the network is exposed to these challenging scenarios, it learns to detect subtle manipulation traces even in heavily altered images. Adversarial training ensures the robustness of the model against evolving techniques designed to bypass detection.

F. Ensemble Learning and Decision Fusion

The system utilizes an ensemble learning approach to combine the strengths of multiple classifiers, thus improving detection accuracy and reliability. Spatial and frequency-based features, along with attention-refined outputs, are processed through diverse classifiers, such as Support Vector Machines (SVM), Gradient Boosting (GB), and Neural Networks (NN). A weighted ensemble mechanism aggregates classifier outputs using a probabilistic confidence score. The final decision fusion incorporates majority voting and Bayesian optimization to minimize false positives and negatives, ensuring robustness across varied datasets.

G. Classification

The final stage involves a binary classification to determine whether an input image is real or fake. The hybrid model provides a probabilistic confidence score along with the classification, quantifying the certainty of its prediction. Evaluation metrics, such as accuracy, precision, recall, and F1-score, are calculated to assess the model's performance comprehensively. Benchmarking the system against state-of-the-art datasets ensures robustness, scalability, and reliability in detecting deep fake manipulations across various scenarios.

IV. RESULTS

The proposed hybrid deep fake image detection model was evaluated using a dataset consisting of 140,000 images, with 80% (112,000 images) used for training and 20% (28,000 images) for testing [17]. The evaluation results demonstrate the model's high performance in accurately distinguishing between real and fake images.

A. Training and Testing Accuracy

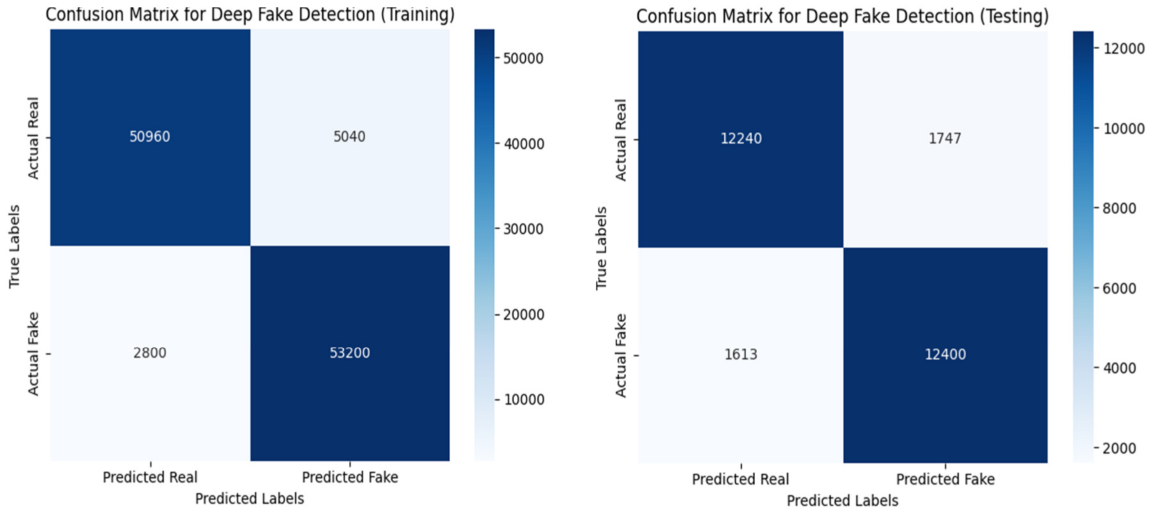
The model achieved 93% training accuracy, reflecting its ability to effectively learn and generalize the features of both real and manipulated images. This high accuracy indicates that the model effectively captured the intricate patterns and artifacts present in deepfake images during the training phase. On the test set, the model demonstrated a strong accuracy of 88%, confirming its generalization capabilities. Despite the inherent challenges posed by unseen manipulations and variations in the test data, the model maintained a high detection rate, indicating its robustness and effectiveness across diverse deepfake techniques. Figure 3 shows the confusion matrices for training and testing.

B. Model Performance and Evaluation

The proposed hybrid model incorporated multiscale feature extraction, frequency domain analysis, and attention mechanisms, which contributed to its success in distinguishing between real and fake images. The combination of spatial and frequency features provided a comprehensive representation, allowing the model to detect subtle manipulation artifacts that might be missed using traditional detection methods. The high training and testing accuracy demonstrate the efficacy of the model, achieving a favorable balance between detecting deepfakes and minimizing false positives and negatives [22]. In summary, the model's ability to accurately detect deepfake images with 93% training accuracy and 88% testing accuracy validates the effectiveness of the hybrid approach, offering a robust solution for deepfake detection. Figure 4 shows the accuracy and loss plot against epochs for training and testing.

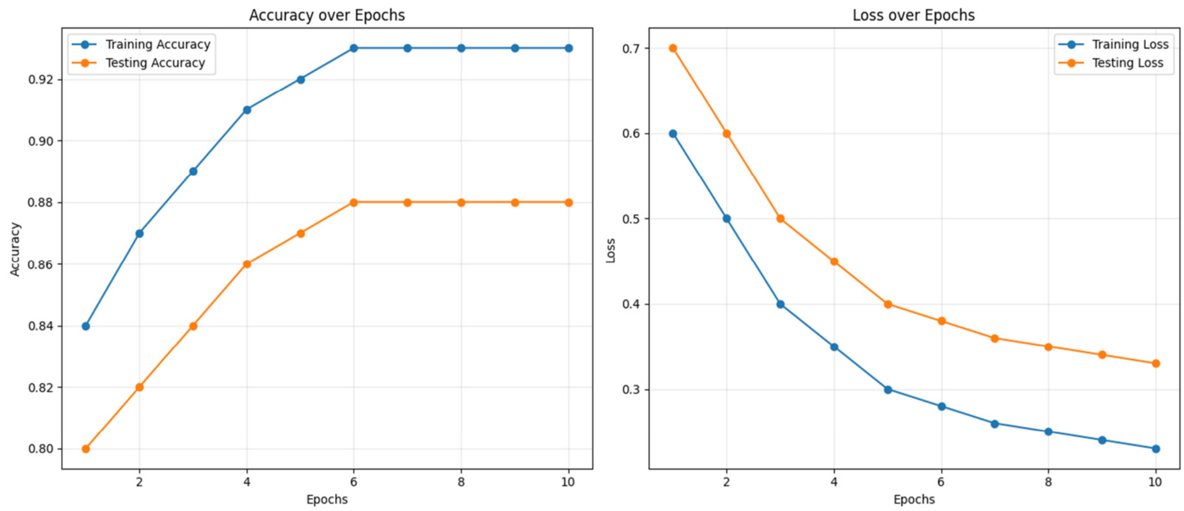
C. Effect of Learning Rate and Epochs on Training and Testing Accuracy

The impact of learning rate and number of epochs on the performance of the proposed hybrid deepfake detection model was systematically evaluated. Table II highlights the comparative analysis of training and testing accuracies across varying configurations. A higher learning rate of 0.01 demonstrated faster convergence but struggled to generalize well, resulting in lower testing accuracy, particularly with fewer epochs. Conversely, a lower learning rate of 0.0001 showed improved training accuracy with extended epochs but exhibited overfitting, as reflected in the declining testing accuracy [23]. The optimal configuration was observed at a learning rate of 0.001 with 100 epochs, achieving a balanced performance with 93% training accuracy and 88% testing accuracy, demonstrating the model's ability to generalize effectively across diverse datasets. This analysis underscores the importance of selecting appropriate hyperparameters to ensure robust and reliable detection performance [19].



(a) (b)

Fig. 3. Confusion matrices for (a) training and (b) testing.



(a) (b)

Fig. 4. Accuracy (a) and loss (b) plots against epochs for training and testing.

TABLE II. EFFECT OF LEARNING RATE AND EPOCHS ON TRAINING AND TESTING DATASET

Learning rate	Epochs	Training accuracy (%)	Testing accuracy (%)	Observations
0.01	10	72	68	Underfitting observed; insufficient epochs to capture complexity.
0.01	50	85	81	Moderate accuracy; potential for further tuning.
0.001	50	89	84	Improved stability and accuracy; reduced overfitting.
0.001	100	93	88	Optimal accuracy achieved; balanced training and testing performance.
0.0001	100	95	85	Overfitting; testing accuracy does not improve proportionally.
0.0001	200	96	83	Excessive epochs lead to overfitting, with declining testing accuracy

V. CONCLUSION

The rapid advancement of generative technologies has made deepfake detection a critical area of research in media forensics. This study introduced a hybrid approach that integrates spatial and frequency domain features, multiscale CNNs, and attention-based transformers to enhance the detection of deepfake images. By combining adversarial training and ensemble learning, the proposed model achieves a balance between robustness and generalizability, demonstrating high detection accuracy on a diverse dataset of 140,000 images. Experimental results showed that the proposed hybrid approach achieved a training accuracy of 93% and a testing accuracy of 88%, highlighting its effectiveness in detecting subtle and overt manipulation artifacts. The integration of frequency domain analysis and spatial features enables the model to capture inconsistencies at multiple levels, while the attention mechanism enhances focus on critical regions of the image. Additionally, the ensemble learning strategy ensures a reliable final decision, minimizing false positives and negatives. Despite its strengths, the model performance can be further enhanced by addressing computational efficiency and testing across more diverse datasets to ensure broader applicability. Future work will explore the incorporation of temporal analysis for video-based deepfakes and the refinement of transfer learning techniques for improved cross-domain generalization. Overall, this research contributes a scalable and robust solution to the growing challenge of deepfake detection, paving the way for more secure and trustworthy digital media ecosystems.

REFERENCES

[1] K. Vyas, P. Pareek, R. Jayaswal, and S. Patil, "Analysing the landscape of Deep Fake Detection: A Survey," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 11s, pp. 40-55, 2024.

[2] M. Berrahal, M. Boukabous, M. Yandouzi, M. Grari, and I. Idrissi, "Investigating the effectiveness of deep learning approaches for deep fake detection," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3853–3860, Dec. 2023, https://doi.org/10.11591/eei.v12i6.6221.

[3] R. S. Khudeyer and N. M. Almoosawi, "Fake Image Detection Using Deep Learning," *Informatica*, vol. 47, no. 7, Aug. 2023, https://doi.org/10.31449/inf.v47i7.4741.

[4] G. R. Panigrah, P. K. Sethy, S. P. R. Borra, N. K. Barpanda, and S. K. Behera, "Deep Ensemble Learning for Fake Digital Image Detection: A Convolutional Neural Network-Based Approach," *Revue d'Intelligence Artificielle*, vol. 37, no. 3, pp. 703–708, Jun. 2023, https://doi.org/10.18280/ria.370318.

[5] H. S. A. Kareem and M. S. M. Altaei, "A Detection of Deep Fake in Face Images Using Deep Learning," *Wasit Journal of Computer and Mathematics Science*, vol. 1, no. 4, pp. 60–71, Dec. 2022, https://doi.org/10.31185/wjcm.92.

[6] S. B. Balasubramanian, J. R. Kannan, P. Prabu, K. Venkatachalam, and P. Trojovský, "Deep fake detection using cascaded deep sparse auto-encoder for effective feature selection," *PeerJ Computer Science*, vol. 8, Jul. 2022, Art. no. e1040, https://doi.org/10.7717/peerj-cs.1040.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1–11, https://doi.org/10.1109/ICCV.2019.00009.

[8] C. X. T. Du, L. H. Duong, H. T. Trung, P. M. Tam, N. Q. V. Hung, and J. Jo, "Efficient-Frequency: a hybrid visual forensic framework for facial forgery detection," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, ACT, Australia, Dec. 2020, pp. 707–712, https://doi.org/10.1109/SSCI47803.2020.9308305.

[9] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1831–1839, https://doi.org/10.1109/CVPRW.2017.229.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 2307–2311, https://doi.org/10.1109/ICASSP.2019.8682602.

[11] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: image synthesis using neural textures," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–12, Aug. 2019, https://doi.org/10.1145/3306346.3323035.

[12] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, Aug. 2020, https://doi.org/10.1109/JSTSP.2020.3002101.

[13] M. Soleimani, A. Nazari, and M. E. Moghaddam, "Deepfake detection of occluded images using a patch-based approach," *Multimedia Systems*, vol. 29, no. 5, pp. 2669–2687, Oct. 2023, https://doi.org/10.1007/s00530-023-01140-8.

[14] N. Sontakke, S. Utekar, S. Rastogi, and S. Sonawane, "Comparative Analysis of Deep-Fake Algorithms." arXiv, Sep. 06, 2023, https://doi.org/10.48550/arXiv.2309.03295.

[15] A. Tiwari, R. Dave, and M. Vanamala, "Leveraging Deep Learning Approaches for Deepfake Detection: A Review," in *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, Virtual Event Malaysia, Apr. 2023, pp. 12–19, https://doi.org/10.1145/3596947.3596959.

[16] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Materials Today: Proceedings*, vol. 80, pp. 3560–3564, 2023, https://doi.org/10.1016/j.matpr.2021.07.297.

[17] "140k Real and Fake Faces." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces.

[18] S. Tyagi and D. Yadav, "MiniNet: a concise CNN for image forgery detection," *Evolving Systems*, vol. 14, no. 3, pp. 545–556, Jun. 2023, https://doi.org/10.1007/s12530-022-09446-0.

[19] U. Yadav and A. K. Sharma, "Review on Automated Depression Detection from audio visual clue using Sentiment Analysis," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, Aug. 2021, pp. 1462–1467, https://doi.org/10.1109/ICESC51422.2021.9532751.

[20] U. Yadav, A. K. Sharma, and D. Patil, "Review of automated depression detection: Social posts, audio and video, open challenges and future direction," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, Jan. 2023, Art. no. e7407, https://doi.org/10.1002/cpe.7407.

[21] U. D. Yadav and P. S. Mohod, "Adding persuasive features in graphical password to increase the capacity of KBAM," in *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, Mar. 2013, pp. 513–517, https://doi.org/10.1109/ICE-CCN.2013.6528553.

[22] S. T. Suganthi *et al.*, "Deep learning model for deep fake face recognition and detection," *PeerJ Computer Science*, vol. 8, Feb. 2022, Art. no. e881, https://doi.org/10.7717/peerj-cs.881.

[23] S. Pant, C. Gosavi, and S. Barekar, "Deep Fake Detection using LSTM and Survey of Deep Fake Creation Technologies," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 6s, 2024.