International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Detecting Deepfakes: Exploring Machine Learning Models for Audio, Video, and Image Analysis

Nilakshi Jain[1], Shwetambari Borade[2], Bhavesh Patel[3], Vineet Kumar[4], Mustansir Godhrawala[5], Shubham Kolaskar[6], Yash Nagare[7], Pratham Shah[8], Jayan Shah[9]

Submitted: 13/03/2024 **Revised**: 24/04/2024 **Accepted**: 01/05/2024

Abstract: The rapid evolution of deepfake technology has created substantial hurdles for the detection of altered media. This study investigates the field of deepfake detection with an emphasis on the use of machine learning techniques in the fields of image, video, and audio analysis. The effectiveness of several machine learning models—Random Forests, Gradient Boosting Machines, Support Vector Machines, Neural Networks, and Convolutional Neural Networks, among others—in identifying deepfakes is compared and contrasted. The analysis outlines the benefits and drawbacks of each model and offers performance insights derived from real-world case studies and research findings. The paper also addresses recent developments in deepfake detection techniques, including ensemble learning approaches and ResNet topologies, which present interesting directions for further research and development in the fight against the spread of manipulated media.

Keywords: *Deepfake Detection, Deepfake Technology Evolution, Ensemble Learning, Machine Learning Techniques, Performance Analysis*

1. Introduction

Machine learning techniques have advanced at a quick pace, revolutionizing several fields, including media manipulation, where the advent of deepfake technology has presented notable hurdles. Deepfake is the term for the artful, frequently malevolent manipulation of audio and visual content using deep learning algorithms. Strong detection techniques are desperately needed to counteract the negative impacts of deepfake content on society, since it is becoming more and more prevalent on the internet. In order to counteract the spread of manipulated media, we apply machine learning methods to the field of deepfake identification in this work.

Deepfake video and audio detection is a challenging task because of the intricate nature of producing extremely realistic forgeries. Because conventional techniques frequently fail to differentiate between authentic and altered content, improved machine learning techniques are being investigated. Researchers have made great progress in creating detection systems that can recognize minute artifacts suggestive of deepfake manipulation by utilizing deep neural networks. These algorithms distinguish between real and fake media by examining a variety of characteristics, including speech patterns, face expressions, and audiovisual discrepancies.

Furthermore, deepfake technology has potential repercussions in fields like identity theft, cybersecurity, and political propaganda, going beyond simple amusement or disinformation. Therefore, it is essential to create trustworthy deepfake detection techniques in order to protect digital media integrity and maintain public confidence in online information. In this work, we provide an overview of the most recent machine learning algorithms used in deepfake detection, emphasizing their benefits, drawbacks, and potential future study areas. By gaining a thorough grasp of these strategies, we hope to support the continuous endeavours to counter the spread of misleading media in the digital era.

1Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0002-6480-2796
2Assistant Professor Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0001-7547-6351
3Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0001-0363-9809
4Founder & Global President, CyberPeace Foundation, Delhi, India
ORCID ID : 0009-0000-3806-7380
5Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0005-4065-4361
6Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0002-1394-7992
7 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0003-1266-3709
8Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0006-0935-6865
9 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0000-9677-9175
** Corresponding Author Email: shwetambari.borade@sakec.ac.in*

2. Literature Survey

Deepfake technology has developed quickly in recent years, making it harder and harder to tell the difference between altered and true media. Therefore, the necessity for efficient and trustworthy techniques to identify deepfakes in audio, video, and picture analysis is increasing. The use of machine learning models has shown promise in solving this issue. The present discourse aims to conduct a comparative analysis of various machine learning models that are utilized in the three domains of audio, video, and picture analysis to detect deepfakes.

Techniques like speech synthesis, voice conversion, and audio splicing can be used to construct audio deepfakes. Machine learning models can be trained on characteristics taken from the audio signal, such as pitch, spectral envelope, and formants, to identify audio deepfakes. To lower the dimensionality of the feature space, feature selection methods like principal component analysis (PCA) and linear discriminant analysis (LDA) can be applied. To distinguish between authentic and fraudulent audio, supervised machine learning methods like random forests and support vector machines (SVM) can be trained on labeled datasets. To find anomalies in the audio signal, unsupervised machine learning techniques like clustering and anomaly detection can be applied. Several successful case studies have demonstrated the effectiveness of machine learning models for audio deepfake detection.

Techniques like lip synchronization, motion transfer, and face swapping can be used to make video deepfakes. Machine learning algorithms are able to assess variables like body motions, face expressions, and inconsistencies in the video frames in order to detect video deepfakes. To distinguish between authentic and false films, supervised machine learning techniques like recurrent neural networks (RNN) and convolutional neural networks (CNN) can be trained on labeled datasets. The generation of synthetic films and the detection of anomalies in the generated frames can also be accomplished with unsupervised machine learning methods like generative adversarial networks (GAN). Machine learning algorithms have proven to be useful for detecting video deepfakes in a number of successful case studies.

Techniques including object removal, face morphing, and image editing can be used to construct image deepfakes. Machine learning algorithms can examine characteristics like texture, color, and form to identify image deepfakes. To find local features in the image, feature extraction methods such accelerated robust feature (SURF) and scale-invariant feature transform (SIFT) can be applied. Using labeled datasets, supervised machine learning algorithms, including k-nearest neighbors (KNN) and decision trees, can be trained to distinguish between real and fraudulent photos. It is also possible to employ unsupervised machine learning techniques like autoencoders and clustering to find abnormalities in the image characteristics. Machine learning algorithms have proven to be useful at detecting image deepfakes in a number of successful case studies.

Robust identification methods are required to stop the spread of deepfake news and stop worldwide threats. Through the application of state-of-the-art Machine Learning (ML) and Deep Learning (DL) techniques, this work proposes a robust deepfake picture identification system. Real-time deepfake detection is made possible by the [1] method, which uses ResNet18's feature vector and SVM classifier to attain an accuracy of 89.5%. Convolutional Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) are used for classification, Error Level Analysis (ELA) is used for pixel-level modification detection, and neural networks (CNNs) are employed for feature extraction. In order to strengthen the resilience of the model, further research will look into other CNN architectures on video datasets and obtain real-world deepfake datasets. This innovative approach empowers a more astute public against potential fake victimization by enabling quick image authenticity assessment.

This [2] analysis explores the evolving landscape of deepfake technology and its connection with image forensics. It focuses on sophisticated machine learning techniques such as CNNs, GANs, and autoencoders, exposing both important obstacles and encouraging progress. Although the accuracies of these techniques are outstanding, ethical considerations highlight the necessity for reliable detection tools. These techniques are useful for recognizing picture modifications in addition to detection. Evaluation techniques emphasize the significance of a comprehensive approach to deepfake difficulties by combining subjective judgments and objective evaluations. This report highlights the progress made in addressing the challenges posed by synthetic media and offers a thorough review of deepfake detection. Continual cooperation, ingenuity, and multidisciplinary methods are critical in the fight against deepfake usage.

This paper [3] proposes a neural network-based method to classify videos, distinguishing between Deepfake and original content with high confidence. Following a review of current algorithms, the design of the project is presented, along with the context and reasoning behind its use. ResNext CNN is used for frame-level feature detection during the model's training on the Celeb-DF dataset. The provided performance results show an average accuracy of 91% when comparing video frames using LSTM. In an effort to slow the transmission of false information via altered digital media, the model can be included into a mobile app to authenticate media material offline, in light of the alarming surge in fake news on social media.

This paper [4] introduces the Celeb-DF dataset, designed for advancing DeepFake detection methods by addressing visual quality disparities between existing datasets and real-world DeepFake videos. By means of Celeb-DF performance evaluations, it highlights the necessity for enhancements in the existing detection techniques. Future work will focus on growing the dataset and improving the structure and efficiency of the synthesis algorithm to produce higher-quality videos. Additionally, we suggest integrating such techniques into the Celeb-DF dataset in order to predict and prevent future anti-forensic strategies utilized by forgers.

This paper [5] focuses on the blurring and transformations required to align generated faces with original videos to present a method for identifying DeepFakes. In order to find discrepancies between the Region of Interest (ROI) and the remainder of the image, artifacts are identified using the Haar Wavelet transformation. The method's efficiency is demonstrated through experimental testing on DeepFake videos. But since there's no magic bullet, we need reliable methods that can manage common picture processing, even if they could have trade-offs in terms of computational complexity.

A large face swap video dataset was developed [6], in order to counter the growing threat of Deepfakes. This dataset is used to support both detection model training and the DeepFake Detection Challenge (DFDC) Kaggle competition. More than 3,400 actors were employed to build the DFDC dataset, which comprises over 100,000 clips produced using a variety of Deepfake, GAN-based, and non-learned techniques. Models trained exclusively on the DFDC dataset are capable of effectively generalizing to real-world Deepfake films, despite the intricacy of Deepfake detection. These models are useful resources for deciphering possibly altered videos.

When celebrities are involved, deepfakes—manipulated videos that employ deep learning to swap faces or create fictitious scenarios—often cause political unrest and disinformation. It is essential to identify and remove these films from social media using easily accessible techniques and increasing quality. This project [7] employs Convolutional Neural Networks and Vision Transformer models to classify manipulated videos, validated on a large dataset from a recent Kaggle challenge.

Reliable detection techniques to alert users to potentially misleading information have been prompted by the rise in deepfake videos. But even with these improvements, these systems' accuracy is still quite low and frequently skewed in favor of the training dataset.. This research [8] analyzes how various training strategies and data augmentation techniques influence CNN-based deepfake detectors, both within the same dataset and across different datasets.

A suggestion is made [9] for a deep convolutional Transformer that combines features from the image, both local and global. Re-attention and convolutional pooling are used to improve these attributes. In order to enhance performance, picture keyframes—which are frequently disregarded—are also used in model training. A visual representation is provided of the feature quantity difference caused by video compression between key and normal image frames. Tested on multiple Deepfake benchmark datasets, the method consistently beats various state-of-the-art techniques in both within- and cross-dataset testing.

3. Background & Evolution of Deepfake Technology

Over the past ten years, deepfake technology has advanced quickly, bringing in a new era where realistic fake films and images can be created with never-before-seen ease and sophistication. Combining the terms "deep learning" and "fake," "deepfake" refers to the process of creating or modifying media content using deep neural networks. Deepfake technology was first tested in research labs, but now that robust computer resources and open-source machine learning frameworks are widely available, anyone with even a basic understanding of computer science can produce realistic-looking fake content.

The creation of generative adversarial networks (GANs), a class of machine learning models first presented by Ian Goodfellow and others in 2014, is credited with spearheading the evolution of deepfake technology. A generator and a discriminator are the two neural networks that make up a GAN. They operate within a game-theoretic framework, with the generator learning to generate realistic outputs that trick the discriminator. Deepfake technology is made possible by GANs producing high-quality synthetic media through this adversarial training process.

Deepfake technology has expanded over time to include a variety of content, such as political propaganda, pornography, and disinformation operations, in addition to its original use in the production of phony celebrity films. With the availability of large-scale datasets and powerful GPUs, along with advancements in deep learning algorithms, deepfake material is becoming more and more believable. As deepfake technology spreads quickly, worries about its possible misuse and social repercussions have grown. As a result, industry stakeholders, legislators, and researchers are investigating ways to limit the negative effects of deepfake technology, such as countermeasures and detection techniques.

4. Machine Learning Models for Deepfake Detection

The incorporation of several machine learning algorithms has led to the advancement of deepfake technology. Popular ensemble learning techniques like Random Forests have demonstrated promise in identifying deepfake content. Random Forests are an efficient way to find anomalies and inconsistencies that point to video manipulation by combining the predictions of several decision trees. However, the dataset's complexity and the hyperparameter selection may have an impact on how well they perform.

Gradient Boosting Machines (GBM) present a potent alternative for detecting deepfakes. By building a series of weak learners one after the other and fixing each other's mistakes, GBM eventually produces a powerful prediction model [10]. Because of this iterative process, GBM is able to achieve high accuracy in identifying real from modified media and capture intricate correlations within the data. However, when dealing with large-scale datasets, GBM could present computational difficulties and necessitate meticulous adjustment to maximize efficiency.

Because Support Vector Machines (SVM) [11] are resilient and interpretable, they have been widely used in deepfake detection. SVM works by locating the best hyperplane in a high-dimensional feature space to divide several classes. Using manually created features or derived representations from pre-trained models, SVM is able to distinguish between real and fake films with high accuracy. SVM may, however, become less effective when dealing with extremely complex datasets and may find it difficult to identify the complex patterns included in deepfake content.

Neural Networks [12], including Multi-layer Perceptrons (MLPs), offer a versatile approach to deepfake detection, leveraging the power of interconnected layers of neurons to learn complex patterns and relationships within data. MLPs excel in capturing nonlinear relationships, making them well-suited for discerning subtle manipulations present in deepfake videos. Nevertheless, the training of MLPs may require large amounts of annotated data and extensive computational resources to achieve optimal performance.

Convolutional Neural Networks (CNNs) [13] have emerged as a cornerstone in deepfake detection, particularly in analyzing visual data. CNNs excel in capturing spatial dependencies within images or video frames, enabling them to identify anomalous artifacts and inconsistencies indicative of deepfake manipulation. Through convolutional layers, pooling operations, and non-linear activation functions, CNNs can effectively discern between authentic and manipulated content with high accuracy. Despite their computational complexity, CNN-based models remain at the forefront of deepfake detection research, driving innovation and advancements in the field.

Deepfake technologies have seen significant advancements with the integration of various machine learning models, each offering unique capabilities in manipulating and detecting manipulated media. Recurrent Neural Networks (RNNs) have been particularly influential in this domain. With their ability to capture temporal dependencies in sequential data, RNNs are well-suited for analyzing video frames over time. This makes them effective in detecting subtle temporal inconsistencies indicative of deepfake manipulation, such as unnatural facial expressions or lip movements. However, RNNs may face challenges in capturing long-term dependencies and may suffer from vanishing or exploding gradient problems during training.

Long Short-Term Memory (LSTM) networks [14], a specialized variant of RNNs, offer a solution to the challenges posed by traditional RNNs. LSTM networks incorporate memory cells that allow them to retain information over long sequences, making them highly effective in analyzing and detecting deepfake content. By selectively retaining and updating information, LSTMs can capture complex temporal patterns present in videos, enabling them to discern between authentic and manipulated media with high accuracy. Nevertheless, LSTMs may require significant computational resources for training and may be prone to overfitting, particularly with limited training data.

Deepfake technology has undergone a radical transformation because to Generative Adversarial Networks (GANs) [15] which allow for the realistic synthesis of pictures and videos. Two neural networks—a discriminator and a generator—play a competitive game together to form a GAN. The discriminator separates authentic media from counterfeit, while the generator creates realistic-looking images or movies. This adversarial training process results in the generation of hyper-realistic deepfake content that is often indistinguishable from genuine media. However, the proliferation of GAN-based deepfakes poses significant challenges for detection methods, as they exploit vulnerabilities in traditional forensic techniques.

Autoencoders [16] make use of the principles of unsupervised learning to provide a novel method for detecting deepfakes. An encoder network compresses input data into a low-dimensional latent space, while a decoder network uses the latent representation to reconstruct the original data. This is how autoencoders function. By reconstructing input data, autoencoders can identify anomalous patterns or artifacts introduced during the deepfake generation process. However, autoencoders may struggle to capture subtle manipulations in highly complex media and may require extensive training on diverse datasets to achieve robust performance.

Gaussian Mixture Models (GMMs) [17] provide a probabilistic framework for deepfake detection, enabling the modeling of complex data distributions. GMMs represent data as a mixture of several Gaussian distributions, each characterized by mean and covariance parameters. By fitting GMMs to feature representations extracted from images or videos, researchers can identify anomalies indicative of deepfake manipulation. However, GMMs may be limited by their assumption of Gaussianity and may struggle to capture non-linear relationships present in high-dimensional data. Additionally, GMMs may require careful parameter tuning and regularization to prevent overfitting and achieve optimal performance.

Deepfake technologies have seen remarkable progress with the integration of diverse machine learning models, each offering unique capabilities in both generating and detecting manipulated media. Hidden Markov Models (HMMs) [18], a probabilistic graphical model, have been employed in analyzing sequential data such as videos for deepfake detection. By modeling the temporal dependencies between consecutive frames, HMMs can capture patterns indicative of deepfake manipulation, such as unnatural transitions or inconsistencies. However, HMMs may face challenges in modeling complex interactions and long-term dependencies present in real-world data, limiting their effectiveness in detecting sophisticated deepfake content.

For deepfake detection tasks, Extreme Gradient Boosting (XGBoost) [19] has become a potent machine learning approach. XGBoost builds a series of decision trees one after the other, with each tree being trained to fix the mistakes of the one before it. XGBoost can detect intricate links in the data and distinguish between real and fake media with great accuracy thanks to this iterative process. To maximize detection performance, however, rigorous adjustment and validation are required because the selection of hyperparameters and the caliber of the training data might have an impact on XGBoost's effectiveness.

CatBoost, a gradient boosting algorithm developed by Yandex, offers a robust and efficient solution for deepfake detection tasks. CatBoost incorporates several innovative features, such as categorical feature handling and robustness to overfitting, making it well-suited for analyzing heterogeneous data characteristic of deepfake content. By leveraging these features, CatBoost can effectively identify anomalies and inconsistencies indicative of manipulation within videos. Moreover, CatBoost's efficient implementation enables fast training and inference, making it suitable for real-time deepfake detection applications.

LightGBM, another gradient boosting algorithm, has gained traction in the field of deepfake detection due to its superior performance and scalability. LightGBM utilizes a novel tree-growing algorithm and histogram-based techniques to achieve faster training times and lower memory consumption compared to traditional gradient boosting methods. This efficiency makes LightGBM well-suited for processing large-scale datasets commonly encountered in deepfake detection tasks. By leveraging LightGBM's capabilities, researchers can develop robust detection systems capable of identifying manipulated media with high accuracy and efficiency.

5. Selection of the model

Determining the best model for deepfake detection depends on various factors such as dataset characteristics, computational resources, and performance metrics. Each model has its strengths and weaknesses, and the suitability of a particular model may vary depending on the specific requirements of the detection task. However, considering the complexity of deepfake manipulation and the need for robust and accurate detection, Convolutional Neural Networks (CNNs) emerge as the most promising choice. Below, are detailed reasons:

Convolutional Neural Networks (CNNs) have demonstrated remarkable success in various computer vision tasks, including deepfake detection. CNNs excel in capturing spatial dependencies within images or video frames, enabling them to identify anomalous artifacts and inconsistencies indicative of deepfake manipulation. For example, CNN-based models have been shown to effectively detect deepfake videos by analyzing subtle discrepancies in facial expressions, skin texture, and lighting conditions.

Evidence from research studies supports the superiority of CNNs in deepfake detection. For instance, in a study conducted by [20], CNN-based models achieved high accuracy in discriminating between genuine and manipulated images, outperforming traditional machine learning algorithms. Similarly, [21] demonstrated the effectiveness of CNNs in detecting deepfake videos by analyzing temporal patterns and spatial features.

Moreover, CNNs offer scalability and flexibility, allowing for the incorporation of additional layers and optimization techniques to improve detection performance. Transfer learning, where pre-trained CNN models are fine-tuned on deepfake detection tasks, further enhances the model's ability to generalize across different datasets and scenarios. This adaptability makes CNNs well-suited for addressing the evolving nature of deepfake technology and emerging manipulation techniques.

Various variations of Convolutional Neural Network (CNN) models have been developed and utilized for deepfake detection, each offering unique architectures and strategies to enhance detection accuracy and robustness. Some notable variations include:

5.1. VGG (Visual Geometry Group)

VGG is characterized by its deep architecture comprising multiple convolutional layers followed by max-pooling layers. Despite its simplicity, VGG has demonstrated effectiveness in deepfake detection tasks due to its ability to capture hierarchical features in images or video frames. By leveraging the depth of the network, VGG can extract complex spatial features indicative of deepfake manipulation.

5.2. ResNet (Residual Network)

ResNet solves the vanishing gradient issue and makes training deeper networks easier by including residual connections that allow the network to learn residual mappings. Because its design can capture complex information in images or videos, it has been frequently used in deepfake detection. ResNet variations have demonstrated promising results in precisely recognizing altered media, including ResNet-50 and ResNet-101.

5.3. InceptionNet

InceptionNet, also known as GoogLeNet, introduces inception modules that allow for the parallel processing of different receptive fields within the same layer. By incorporating multiple convolutional operations in parallel, InceptionNet can capture diverse spatial features at different scales, making it effective in discerning anomalies indicative of deepfake manipulation.

5.4. DenseNet (Densely Connected Convolutional Networks)

DenseNet creates dense connections between layers so that all feature maps from previous layers are fed into the current layer. More effective information propagation and feature extraction are made possible by this dense connection, which also encourages gradient flow throughout the network and makes feature reuse easier. Due to their ability to capture spatial correlations inside pictures or video frames, DenseNet architectures have demonstrated potential in deepfake detection applications.

5.5. MobileNet

Depth-wise separable convolutions are used by MobileNet to minimize model size and computational cost while preserving high accuracy. Particularly well-suited for contexts with limited resources are mobile devices and edge computing platforms. For real-time deepfake detection applications, MobileNet variants like MobileNetV2 and MobileNetV3 provide effective solutions.

5.6. EfficientNet

EfficientNet introduces compound scaling to balance model size and performance across different network depths, widths, and resolutions. By systematically scaling the model's architecture, EfficientNet achieves optimal trade-offs between computational efficiency and detection accuracy. This scalability makes EfficientNet well-suited for deepfake detection tasks requiring robust performance across diverse datasets and scenarios.

Each of these variations of CNN models offers unique advantages and capabilities in deepfake detection, enabling researchers to develop sophisticated detection systems capable of combating the proliferation of manipulated media across digital platforms. The choice of model depends on factors such as dataset characteristics, computational resources, and performance requirements.

Determining which CNN model outperforms the others in deepfake detection requires careful consideration of various factors such as dataset characteristics, computational resources, and evaluation metrics. While each CNN model has its strengths and weaknesses, ResNet stands out as a top performer in many scenarios due to its innovative architecture and proven effectiveness in capturing intricate features indicative of deepfake manipulation.

ResNet, or Residual Network, stands out as a superior model for deepfake detection due to several key factors, supported by evidence from various research studies:

Deep Architecture with Residual Connections: ResNet's innovative architecture incorporates residual connections, allowing the network to learn residual mappings. This approach addresses the vanishing gradient problem encountered in training deep neural networks by facilitating the flow of gradients throughout the network. As a result, ResNet can effectively capture complex spatial features and patterns within images or video frames, enabling it to discern subtle anomalies indicative of deepfake manipulation [22].

State-of-the-Art Performance: Numerous studies have demonstrated ResNet's superior performance in deepfake detection tasks compared to other CNN models. For example, [23] found that ResNet architectures consistently outperformed VGG, InceptionNet, and MobileNet in accurately identifying deepfake videos across different datasets and evaluation metrics. Similarly, research by [23]showcased the effectiveness of ResNet-based models in detecting deepfake content by analyzing spatial features and temporal patterns present in videos.

Transfer Learning Capabilities: ResNet's pre-trained models, such as ResNet-50 and ResNet-101, offer transfer learning capabilities that enhance detection performance. By fine-tuning pre-trained models on deepfake detection tasks with relatively small annotated datasets, researchers can leverage the representations learned from large-scale datasets (e.g., ImageNet) to improve detection accuracy and generalization across different domains [24].

Efficient Feature Extraction: ResNet's residual connections facilitate efficient feature extraction, enabling the model to capture subtle discrepancies and inconsistencies indicative of deepfake manipulation. By leveraging hierarchical representations learned across multiple layers, ResNet can effectively discriminate between genuine and manipulated media with high accuracy, even in the presence of complex variations and transformations [22].

In summary, ResNet's deep architecture with residual connections, state-of-the-art performance, transfer learning capabilities, and efficient feature extraction make it a preferred choice for deepfake detection tasks. While other CNN models may offer unique advantages in specific scenarios, ResNet's consistent performance and robustness position it as a top performer in the field of deepfake detection.

6. Conclusion

In conclusion, in the current digital environment, identifying deepfake content continues to be a significant difficulty. Convolutional Neural Networks (CNNs), in particular, are machine learning models that have become extremely effective at detecting altered media in the audio, video, and image domains. ResNet is a CNN version that is distinguished by its deep architecture with residual connections, cutting-edge performance, capacity for transfer learning, and effective feature extraction. With evidence from much research, ResNet outperforms other CNN models in terms of accuracy when it comes to identifying deepfake content. To address the dynamic nature of deepfake technology and create reliable detection techniques that can lessen its detrimental consequences on society, continued study is necessary. By leveraging innovative techniques and interdisciplinary collaborations, researchers can advance the field of deepfake detection and safeguard the integrity of digital media in the digital age.

Acknowledgements

We extend our sincere gratitude to the CyberPeace Foundation for their generous support and grant provided for our deepfake project. Their commitment to fostering research and innovation in the field of cybersecurity has been invaluable in advancing our efforts to combat the proliferation of manipulated media. We deeply appreciate their dedication to promoting cyberpeace and safeguarding digital integrity.

References

[1] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, “Deep fake detection and classification using error-level analysis and deep learning,” Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-34629-3.

[2] N. Jain et al., “Deepfake Technology and Image Forensics: Advancements, Challenges, and Ethical Implications in Synthetic Media Detection,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 16s, 2024.

[3] V. V. V. N. S. Vamsi et al., “Deepfake detection in digital media forensics,” Global Transitions Proceedings, vol. 3, no. 1, pp. 74–79, Jun. 2022, doi: 10.1016/j.gltp.2022.04.017.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,” Sep. 2019, doi: https://doi.org/10.48550/arXiv.1909.12962.

[5] Zankoya Dihuk, Institute of Electrical and Electronics Engineers, and Institute of Electrical and Electronics Engineers. Iraq Section, Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform. 2020.

[6] B. Dolhansky et al., “The DeepFake Detection Challenge (DFDC) Dataset,” Jun. 2020, doi: https://doi.org/10.48550/arXiv.2006.07397.

[7] A. Seth and A. K. Gogineni, “Detection of Deep-fakes in Videos using CNN and Transformers”, doi: 10.13140/RG.2.2.23238.60480.

[8] L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, “Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection,” in 2020 IEEE International Workshop on Information Forensics and Security, WIFS 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/WIFS49906.2020.9360901.

[9] T. Wang, H. Cheng, K. P. Chow, and L. Nie, “Deep Convolutional Pooling Transformer for Deepfake Detection,” Sep. 2022, doi: 10.1145/3588574.

[10] S. Borade, P. C. Doshi, and D. B. Patel, “MaliceSpotter: Revolutionizing Cyber Security with Machine Learning for Phishing Resilience,” Indian J Sci Technol, vol. 17, no. 10, pp. 870–880, Mar. 2024, doi: 10.17485/IJST/v17i10.148.

[11] M. A. Chandra and S. S. Bedi, “Survey on SVM and their application in image classification,” International Journal of Information Technology (Singapore), vol. 13, no. 5, 2021, doi: 10.1007/s41870-017-0080-1.

[12] R. B. Shobha Rani, P. Kumar Pareek, S. Bharathi, and G. Geetha, “Deepfake Video Detection System Using Deep Neural Networks,” in 2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICICACS57338.2023.10099618.

[13] I. Ilhan, E. Bali, and M. Karakose, “An Improved DeepFake Detection Approach with NASNetLarge CNN,” in 2022 International Conference on Data Analytics for Business and Industry, ICDABI 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 598–602. doi: 10.1109/ICDABI56818.2022.10041558.

[14] B. Chen, T. Li, and W. Ding, “Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM,” Inf Sci (N Y), vol. 601, pp. 58–70, Jul. 2022, doi: 10.1016/j.ins.2022.04.014.

[15] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, “GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection,” IEEE Journal on Selected Topics in Signal Processing, vol. 14, no. 5, 2020, doi: 10.1109/JSTSP.2020.3007250.

[16] M. M. Price and M. A. Price, “Methods and systems for detecting deepfakes,” US Patent 10,929,677, 2021.

[17] C. Zhang, D. Hu, and T. Yang, “Research of artificial

intelligence operations for wind turbines considering anomaly detection, root cause analysis, and incremental training," Reliab Eng Syst Saf, vol. 241, 2024, doi: 10.1016/j.ress.2023.109634.

[18] S. Goumiri, D. Benboudjema, and W. Pieczynski, "A new hybrid model of convolutional neural networks and hidden Markov chains for image classification," Neural Comput Appl, vol. 35, no. 24, 2023, doi: 10.1007/s00521-023-08644-4.

[19] S. Kabiraj et al., "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, pp. 2020–2023, 2020, doi: 10.1109/ICCCNT49239.2020.9225451.

[20] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting Fake News: Image Splice Detection via Learned Self-Consistency," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018. doi: 10.1007/978-3-030-01252-6_7.

[21] Y. Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection," IEEE Access, vol. 11, pp. 22081–22095, 2023, doi: 10.1109/ACCESS.2023.3251417.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016. doi: 10.1109/CVPR.2016.90.

[23] J. Gadgilwar, K. Rahangdale, O. Jaiswal, P. Asare, P. Adekar, and Prof. L. Bitla, "Exploring Deepfakes - Creation Techniques, Detection Strategies, and Emerging Challenges: A Survey," Int J Res Appl Sci Eng Technol, vol. 11, no. 3, 2023, doi: 10.22214/ijraset.2023.49681.

[24] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019. doi: 10.1109/CVPR.2019.00277.