Deep Fusion: A Robust Framework for Deepfake Video Classification Using Convolutional and Recurrent Neural Networks

Shaista Sultana [1], Dr M Swapna [2]
Stanley College of Engineering and Technology for Women [1,2]

Abstract **- Deepfake videos, developed using sophisticated artificial intelligence methods, challenge the credibility and security of digital content by fabricating highly realistic but deceptive visuals. To counter this, a hybrid deep learning framework has been implemented, combining InceptionV3 for capturing detailed spatial features from video frames with Gated Recurrent Units (GRUs) for recognizing sequential temporal patterns. The approach involves preprocessing video data, extracting intricate frame-level features, and analyzing temporal consistencies through GRUs to identify manipulations effectively. Binary cross- entropy loss directs the training process, with early stopping mechanisms ensuring robust and efficient learning. Tested on a curated dataset, this method provides a reliable solution for preserving the authenticity of multimedia content and addressing the risks posed by deepfake technologies.**

Key Words: **Deepfake Detection, InceptionV3, GRU, Temporal Dependencies, Feature Extraction, Video Classification, Deep Learning, Multimedia Content Integrity, AI Algorithms, Convolutional Neural Networks, Recurrent Neural Networks, Early Stopping, Binary Cross-Entropy, Digital Security, Media Forensics.**

1. INTRODUCTION

Deepfake videos, generated using advanced AI algorithms, modify existing video content to depict individuals performing actions or making statements they never did. These videos represent substantial risks to privacy, security, and the spread of disinformation. Traditional methods for detecting deepfakes relied on manual expert analysis, examining factors like video quality inconsistencies, lighting artifacts, and physiological anomalies such as unnatural eye movements. However, such techniques often proved inefficient against sophisticated deepfakes and were time-consuming. Recent advancements employ deep learning models to enhance detection processes. InceptionV3, as a convolutional neural network (CNN), excels at extracting detailed spatial information from video frames, while Bidirection 1], InceptionV3 utilizes diverse convolutional filters to ensure efficient feature extraction at various scales, enhancing precision and reliability.Moreover, [2] highlights the ability of Bi-GRUs to process data bidirectionally, enabling comprehensive temporal analysis for improved classification of real and fake videos. Together, these models offer robust solutions for detecting deepfake content, significantly outperforming traditional techniques.

1.1 Overview

InceptionV3, a state-of-the-art convolutional neural network (CNN), is highly effective in extracting intricate features from video frames. By utilizing multiple convolutional filters operating at various scales, it efficiently captures essential spatial details. For sequence modeling, Gated Recurrent Units (GRUs) are employed to analyze temporal dependencies within frames. The combination of these models enables the accurate detection and classification of deepfake videos by leveraging both spatial and temporal data. According to [3], InceptionV3's ability to extract diverse features enhances its precision, while [4] highlights GRU's efficiency in capturing temporal patterns, leading to comprehensive analysis. Together, these models contribute to developing robust systems for detecting and classifying deepfakes, supporting the integrity of digital media. As noted in [3], advancements in AI and machine learning continue to combat evolving deepfake techniques. Moreover, [4] emphasizes that these advancements are pivotal in ensuring a safer and more credible digital landscape.

1.2 Problem Statement

The rapid growth of deepfake videos poses critical threats to privacy, security, and public trust in digital media. These artificially manipulated videos create

false narratives, spread misinformation, and damage personal reputations. Effective detection methods are urgently needed to mitigate these impacts. Traditional approaches, such as manual inspection and forensic analysis, are time-intensive and lack scalability. To overcome these limitations, advanced deep learning techniques are increasingly being explored. According to [5], these methods leverage convolutional and recurrent neural networks to improve the efficiency and accuracy of deepfake detection.

1.3 Aim & Scope

Aim: This research aims to develop a machine learning-based system for identifying deepfake videos using InceptionV3 and GRU models. The primary objective is to develop a dependable model that can manage many datasets, efficiently evaluate temporal and spatial features in video data, and minimize false positives and negatives. The system uses advanced modeling techniques to increase media security and reliability. This tactic is crucial for lowering the risks associated with deepfakes and fostering trust in digital platforms.

Scope: The project is centered on creating a machine learning system that incorporates InceptionV3 and GRU models for detecting deepfake content. This includes working with datasets containing diverse features, such as frame sequences and facial landmarks. Techniques addressing class imbalance will be implemented to improve prediction accuracy. As highlighted in [7], performance metrics like accuracy, precision, recall, and F1-score will guide evaluation to ensure reliable detection. The research aims to contribute to secure, trustworthy digital media by offering a scalable and efficient detection solution suitable for real-time applications.

1.4 Objectives:

This project strives to develop an advanced deep learning system for accurately detecting and classifying deepfake videos by combining InceptionV3 for spatial feature extraction and GRU for temporal modeling. With the proliferation of deepfake content, innovative methods are essential to preserve digital media integrity. Key objectives include:

1. Develop a deep learning-based model for precise detection of deepfake videos.
2. Extract high-level visual features from video frames.
3. Analyze temporal dependencies in video sequences.
4. Ensure accurate differentiation between authentic and deepfake videos.

As [8] underscores, combining convolutional and recurrent neural networks offers robust solutions for addressing challenges in deepfake detection. By achieving these objectives, the project supports combating misinformation and safeguarding digital media credibility in the modern era.

2. LITERATURE REVIEW

A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, 2024. Deepfake detection using deep learning methods: a systematic and comprehensive review. This review systematically evaluates existing deep learning methodologies for detecting deepfakes, focusing on the performance of Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and hybrid approaches. It also explores benchmark datasets such as Celeb-DF and DeepFake Detection Challenge Dataset (DFDC). The authors highlight key challenges, such as adversarial attacks on detection systems and the need for more diverse datasets to enhance model generalizability. [1]

H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, and O. Etzioni, 2024. The tug-of- war between deepfake generation and detection. This paper explores the dynamic progression of deepfake generation and detection techniques. The authors provide an analysis of popular algorithms, including those powered by GANs, and evaluate the trade- offs between quality of fake media and detection accuracy. They suggest incorporating explainable AI for enhancing trust in detection systems and propose game-theoretical models to anticipate future developments in the deepfake arms race. [2]

J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, 2023. An enhanced deep learning-based deepfake video detection and classification system. The study introduces an innovative system that employs hybrid architectures combining CNNs and Transformer models. Special emphasis is placed on leveraging temporal features in videos to improve classification. Additionally, the framework integrates a confidence scoring mechanism to reduce false-positive rates in real-world applications like law enforcement and media verification. [3]

P. Sugavaneshwari, R. Sreelekha, M. Sathya Jothi, S. Swetha, and J. Rudhra, 2024. Deepfake detection using deep learning. The authors present an exploration of advanced preprocessing techniques such as frequency-domain analysis and the role of synthetic training data. They discuss how techniques like Variational Autoencoders (VAEs) complement CNN-based detection systems. Furthermore, the paper addresses the computational challenges and proposes optimized solutions for real-time deployment on edge devices. [4]

N. Rathoure, R. K. Pateriya, N. Bharot, and P. Verma, 2024. Combating deepfakes: a comprehensive multilayer deepfake video detection framework. *Multimedia Tools and Applications.* This multilayer framework leverages feature extraction at multiple stages, including pixel-level analysis, temporal consistency checks, and deep neural network- based feature fusion. The authors provide a comparative study of detection accuracy across video quality levels and compression artifacts, showing significant robustness in low-quality data environments. [5]

M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, 2024. Deepfake detection: a systematic literature review. The review categorizes existing methods into detection strategies using spatial, temporal, and frequency-domain features. It also addresses critical limitations, such as scalability issues and biases in current detection models. The authors advocate for a standardized evaluation framework to unify research efforts and improve reproducibility in deepfake detection studies. [6]

A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, 2024. Deepfake video detection: challenges and opportunities. This paper outlines societal and technical challenges in combating deepfake threats, emphasizing the role of interdisciplinary collaboration between computer scientists, policymakers, and ethicists. The authors discuss the ethical implications of detection tools, particularly regarding privacy and surveillance, and propose a roadmap for the next decade in deepfake research. [7]

M. M. Khan and K. N. Alam, 2022. Detecting deepfake images using deep learning techniques and explainable AI methods. *Intelligent Automation & Soft Computing*, vol. 35, no. 2, July. This study evaluates the interpretability of deep learning-based methods for detecting manipulated facial images. By integrating explainable AI techniques, such as SHAP (Shapley Additive Explanations), the work seeks to bridge the gap between model complexity and user trust. The authors also explore the integration of attention mechanisms to enhance detection performance. [8]

S. Ahmed Khan and D.-T. Dang-Nguyen, 2023. Deepfake detection: a comparative analysis. *SFI-MediaFutures, University of Bergen, Norway.* This comprehensive analysis compares conventional and state-of-the-art deepfake detection approaches, including heuristic methods and advanced neural networks. Performance benchmarks are provided for various datasets, along with a discussion on the trade-offs between detection accuracy, computational cost, and real-time feasibility. [9]

T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, 2024. Deep learning for deepfakes creation and detection: a survey. This survey categorizes deepfake creation and detection methods based on complexity, data requirements, and computational efficiency. The authors discuss advancements in GAN-based detection methods and highlight emerging areas like audio-visual synchronization analysis and adversarial training to improve model robustness against unseen manipulations. [10]

3. EXISTING SYSTEM

The Existing deepfake detection systems face several limitations that hinder their effectiveness. Many systems rely on heuristic-based rules designed by domain experts to identify suspicious activities. These rules can be too rigid and struggle to keep up with the rapidly evolving nature of deepfake technologies. As deepfake techniques become more sophisticated, static detection rules become less effective, necessitating more adaptable solutions. Additionally, handling imbalanced datasets is a significant challenge in deepfake detection.

Datasets often contain far more genuine videos than fake ones, leading to high false positive rates where genuine videos are incorrectly flagged as fake. This imbalance reduces the overall efficiency and reliability of the detection system. Computational complexity is another limitation, especially for systems employing recurrent neural networks (RNNs).

Training and deploying these models require significant computational resources, which may not be feasible for all applications. Furthermore, some models may perform well on the dataset they were trained on but fail to generalize to new, unseen data. This lack of robustness limits their effectiveness in real-world scenarios, where deepfakes can exhibit diverse characteristics and manipulations. Therefore, there is a pressing need for advanced models that can handle these challenges effectively.

4.SYSTEM DESIGN

The deepfake detection challenge dfdc dataset accessible on kaggle is a powerful tool for advancing the creation and evaluation of models designed to identify deepfake videos organized into two primary directoriestrainsamplevideos and testvideosit includes 401 labeled training clips and 400 testing videos providing a strong foundation for the systematic development of detection systems to prepare the dataset individual frames are extracted from each video for detailed analysis advanced computer vision techniques are then applied to detect and crop faces within these frames honing in on areas most likely to exhibit manipulations while discarding irrelevant details the dataset is further refined through preprocessing steps like normalization scaling and augmentation techniques including adjustments in brightness flipping and rotations these enhancements ensure not only consistency but also the diversity required for robust model.

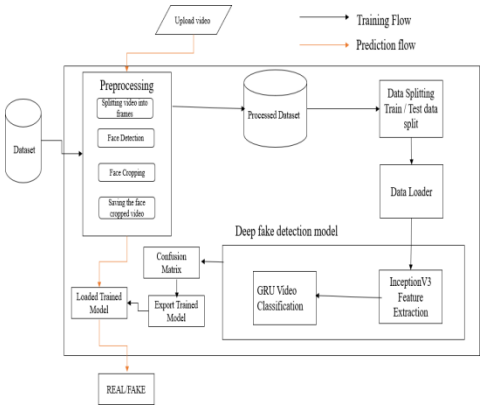

Figure 4.1 System Architecture

training the dataset is divided into training and testing subsets often using ratios such as 8020 or 7030 the testing portion evaluates the models ability to generalize to unseen data while the training subset is used to refine the models predictive performance cross- validation techniques like k-fold cross-validation further bolster reliability by systematically cycling through subsets to test different parts of the model a hybrid model architecture is employed integrating spatial and temporal analysis for a comprehensive approach spatial features are extracted from video frames using inception v3 a convolutional neural network cnn known for its efficiency and precision in identifying complex patterns meanwhile gated recurrent units grus handle sequential data effectively analyzing the temporal changes within video sequences this collaboration ensures that the model not only identifies subtle visual artifacts but also tracks their progression over time training begins with data being fed into the model via a data loader which ensures efficient processing early stopping is utilized to prevent overfitting by halting the process when validation loss ceases to improve model checkpoints are implemented to save configurations that demonstrate optimal performance while hyperparameter tuning adjusts factors such as learning rates and batch sizes to fine-tune the training process once trained the models effectiveness is measured using evaluation metrics like precision recall accuracy and confusion matrices these metrics provide insight into how well the model performs and highlight areas requiring further improvement finally the trained system is deployed to classify videos as either real or fake predictions include confidence scores and can also highlight manipulated regions within videos providing actionable information by detecting alterations the system plays a vital role in enhancing security curbing the spread of misinformation and preserving the authenticity of digital media this process transforms raw video data into a functional and efficient tool capable of addressing real-world challenges associated with deepfake technologies.

5.DATA SET INFORMATION

The Deepfake Detection Challenge (DFDC) dataset, hosted on Kaggle, is a comprehensive collection of video clips created to facilitate the development and testing of models capable of identifying deepfake content. Deepfakes, generated through advanced AI techniques, can alter audio or visual media to mimic realistic portrayals of individuals, posing significant ethical concerns. This dataset serves as a crucial resource for researchers aiming to build reliable detection systems.

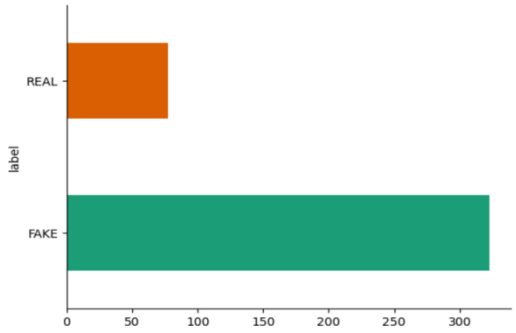

Figure 5.1 Deepfake Detection Challenge (DFDC) dataset

The dataset features a diverse range of video samples designed to ensure models can generalize effectively across various scenarios. It is organized into two primary directories: and , containing 401 and 400 video samples, respectively. This structured organization supports efficient model training and evaluation processes. Each sample is meticulously labeled, and the dataset also includes a metadata file, , which provides detailed information about the video clips.

The metadata file comprises key columns such as , , and . The column categorizes videos as either real or fake, providing essential binary labels for supervised learning models. The column divides the videos into training and testing sets, enabling a systematic approach to model validation. The column links fake videos to their authentic counterparts, creating a traceable connection for further analysis.

This dataset is instrumental in advancing deepfake detection technologies. It encourages experimentation with various methods, such as convolutional neural networks (CNNs) for spatial analysis and recurrent neural networks (RNNs) for temporal feature extraction.
These techniques, along with hybrid approaches, allow models to identify inconsistencies that indicate manipulated content.

The DFDC dataset is well-suited for real-world applications, such as misinformation prevention, enhancing security measures, and verifying media authenticity. Its detailed labeling and comprehensive structure provide a solid foundation for developing scalable models capable of addressing the challenges posed by synthetic media. By offering a diverse and organized resource, the DFDC dataset contributes significantly to ongoing efforts to counter deepfake misuse

6.PROPOSED METHODOLOGY

The proposed methodology The project employs a hybrid architecture combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively detect and classify deepfake videos. The CNN component uses the Inception V3 model, which excels at extracting spatial features and capturing intricate details from video frames. Inception V3's deep and wide architecture makes it particularly effective for analyzing complex visual content, ensuring high accuracy in processing individual video frames. This model builds on the principles introduced by its predecessor, Inception V1, which used parallel layers with different filter sizes to capture diverse spatial features while minimizing overfitting and computational inefficiency

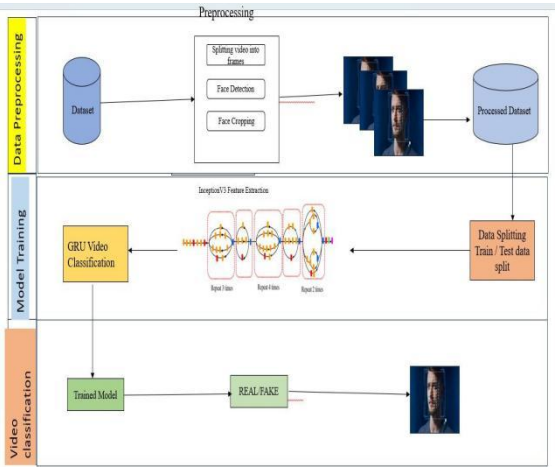

Figure 6.1 Swimlane Diagram

The proposed methodology The project employs a hybrid architecture combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively detect and classify deepfake videos. The CNN component uses the Inception V3 model, which excels at extracting spatial features and capturing intricate details from video frames. Inception V3's deep and wide architecture makes it particularly effective for analyzing complex visual content, ensuring high accuracy in processing individual video frames. This model builds on the principles introduced by its predecessor, Inception V1, which used parallel layers with different filter sizes to capture diverse spatial features while minimizing overfitting and computational inefficiency.

To complement the spatial analysis, the project

incorporates RNNs, specifically Gated Recurrent Units (GRUs) and Bidirectional GRUs (Bi-GRUs), for temporal analysis. GRUs address key challenges of traditional RNNs, such as vanishing and exploding gradients, by employing gating mechanisms like the Update Gate and Reset Gate. These gates enable the efficient learning of long-term dependencies, making GRUs well-suited for identifying subtle temporal inconsistencies that may indicate deepfake content. The Bi- GRU further enhances this capability by processing sequences in both forward and backward directions, enabling a more comprehensive understanding of the temporal dynamics in video data.

This hybrid approach combines the strengths of CNNs and RNNs, resulting in improved detection accuracy and reduced false positives. By leveraging spatial features extracted by the CNN and temporal patterns learned by the RNN, the system ensures a robust and adaptive framework for combating deepfakes. The fusion of Inception V3 with GRUs and Bi-GRUs provides a reliable solution against the evolving threats of deepfake videos, making it suitable for diverse applications, including security and media verification.

The Inception V3 model, which is a member of the Inception family, uses cutting-edge methods to increase accuracy and processing efficiency. Along with 3×3 max pooling layers, its predecessor, Inception V1, introduced the idea of parallel layers with filters of different sizes, including 1×1, 3×3, and 5×5 convolutions. Despite its effectiveness, Inception V1 has computational issues, especially with the 5x5 convolution layer. This was fixed in Inception V3 by prioritizing 1×1 convolutions above other convolutional layers, which greatly reduced dimensions and increased processing performance without sacrificing accuracy.

GRUs became a more straightforward and effective substitute for Long Short-Term Memory (LSTM) networks on the RNN side in 2014. Due to their ability to store crucial information over lengthy sequences and their ease of training, GRUs are especially good in sequential data problems. The drawbacks of conventional RNNs, such as their subpar performance on long-range dependencies, can be addressed by their gating methods.
Furthermore, by conducting both forward and backward data analysis, Bi-GRUs expand the capabilities of GRUs. For applications like natural language processing and video analysis, the model's capacity to capture both past and future context is essential.
GRUs and Bi-GRUs have demonstrated remarkable versatility in a wide range of applications, ranging from time series prediction and speech recognition to video analysis and natural language problems. They offer an excellent combination between simplicity and performance, which makes them a popular choice in the deep learning field. By combining Inception V3, GRUs, and Bi-GRUs, the proposed methodology provides a robust and comprehensive approach for detecting and classifying deepfake videos

7.RESULTS AND DISCUSSION

Using a deep learning model, the system uses a GRU to analyze temporal patterns and InceptionV3 to extract visual features for video classification. The model's confidence in identifying the video as "real" or "fake" is shown by the probability it produces.

Score for Confidence and Threshold:

With a range of 0 to 1, the confidence score indicates how certain the algorithm is. While scores around 0 indicate greater trust in a "real" classification, scores near 1 indicate stronger confidence in a "fake" classification.

- The present implementation determines the final categorization using a threshold of 0.5:
- The video is deemed "fake" if the Confidence Score is more than 0.5.
- Then "Real" is the classification for the video if the Confidence Score is less than 0.5.

Figure 7.1 Video classification Example 1

- A confidence score of 0.41 indicates that the

model is slightly more confident that the video is real.

- This is because the score is below the 0.5 threshold
- values closer to 0 indicate higher confidence in a "real" classification

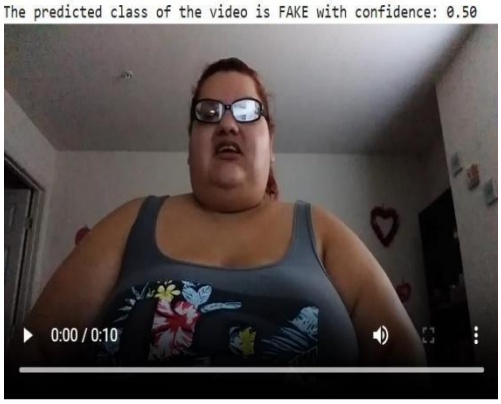

Figure 7.2 Video classification Example 2

- A confidence score of 0.50 indicates the model is slightly more confident that the video is fake.
- However, this confidence level is relatively low, suggesting some uncertainty in the prediction. Nonetheless, based on the established threshold of 0.5, the video would still be classified as "fake."

8.CONCLUSION

The project integrates the complementary strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to form a comprehensive solution for detecting and classifying deepfake videos. The Inception V3 model excels at extracting intricate spatial features from video frames, ensuring high accuracy in capturing complex visual patterns. The use of Gated Recurrent Units (GRUs) and Bidirectional GRUs (Bi-GRUs) enhances the model's ability to analyze temporal dynamics, enabling the identification of subtle inconsistencies that are indicative of deepfakes. By combining these spatial and temporal insights, the hybrid architecture achieves remarkable detection accuracy while minimizing false positives. This framework holds immense potential in addressing the growing threats posed by deepfake technologies. Its adaptability to diverse datasets and evolving techniques makes it a robust tool for combating the misuse of synthetic media. Moreover, the project's focus on efficiency ensures its applicability in real-world scenarios, such as security, media authentication, and online content verification. The methodology's strong foundation paves the way for innovative solutions in safeguarding digital integrity.

9.FUTURE SCOPE

The project offers a promising foundation for future advancements in deepfake detection and classification. Incorporating transfer learning with other state-of-the-art models could further enhance the system's adaptability to unseen data and emerging deepfake techniques. Expanding the dataset to include diverse scenarios, such as different languages, accents, and cultural contexts, would improve the model's generalizability across global applications.

REFERENCES

[1] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review," 2024.

[2] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, and O. Etzioni, "The Tug-of-War Between Deepfake Generation and Detection," 2024.

[3] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, "An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System," 2023.

[4] P. Sugavaneshwari, R. Sreelekha, M. Sathya Jothi, S. Swetha, and J. Rudhra, "Deepfake Detection Using Deep Learning," 2024.

[5] N. Rathoure, R. K. Pateriya, N. Bharot, and P. Verma, "Combating Deepfakes: A Comprehensive Multilayer Deepfake Video Detection Framework," *Multimedia Tools and Applications*, 2024.

[6] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," 2024.

[7] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake Video Detection: Challenges and Opportunities," 2024.

[8] M. M. Khan and K. N. Alam, "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods," *Intelligent Automation & Soft Computing*, vol. 35, no. 2, July 2022.

[9] S. Ahmed Khan and D.-T. Dang-Nguyen, "Deepfake Detection: A Comparative Analysis,"

SFI-MediaFutures, University of Bergen, Norway, 2023.

[10] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep Learning for Deepfakes Creation and Detection: A Survey," 2024.