

Scalable System for Identifying and Flagging Misinformation Using AI/ML and DL

Ashik Shaikh^{1,*}, Sharmad Shringarpure¹, Shiksha Dubey²

Abstract

The fast spread of realistic fake media means we need to make a system to identify false information that can work with a lot of data. This system is called the Scalable Misinformation Identification System or SMIS for short. SMIS helps keep our society and democracy safe from the threats of news on the internet. The old ways of checking for news and simple machine learning models are not good enough to find the small clues in modern fake media that are made by advanced computers. The SMIS uses a Deep Learning framework that uses Transfer Learning to see which of the best Convolutional Neural Networks works the best. These networks are Xception, InceptionV3, ResNet50 and MobileNetV2. They are very good at finding the clues in fake media that are made by computers. The SMIS checks these networks to find the one for finding fake media. The results show that these complex models are much better than the ways of checking for fake news. They can find the clues that show when media has been manipulated. By choosing the model the SMIS can work very fast and can be used on the cloud. This means the SMIS can find media in real time. The SMIS is very important for keeping people trusting the news and making sure the information on the internet is true. The Scalable Misinformation Identification System is a tool for keeping our society safe from fake news, in a world where computers are making more and more of the media we see.

Keywords: Identifying misinformation, deep learning, transfer learning, detecting deepfakes, Xception, image classification.

INTRODUCTION

The integrity of the digital information ecosystem faces major challenges owing to the rapid growth and increasing complexity of synthetic media, such as deepfakes. Advances in Generative Adversarial Networks (GANs) and rendering techniques have made digital manipulation nearly perfect, which threatens public trust and affects digital forensics. Current defense methods often rely on manual

feature engineering or simple convolutional neural network (CNN) models; however, these approaches are proving to be inadequate. They lack the scalability and general ability required to handle the vast amounts of content uploaded daily, especially when dealing with dynamic adversarial attacks. This gap creates a significant weakness: existing AI models are too fragile to ensure trust, and the sheer volume of data makes comprehensive manual verification impossible.

To fill this gap, this study introduces the Scalable Misinformation Identification System (SMIS), an intelligent prototype aimed at providing automated and reliable deepfake detection. A key part of this study is the systematic

*Author for Correspondence

P.B. Kudal
E-mail: shaikhashik0330@gmail.com

¹Research Scholar, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India

²Assistant Professor, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India

Received Date: March 10, 2026

Accepted Date: June 26, 2026

Published Date: June 30, 2026

Citation: Ashik Shaikh, Sharmad Shringarpure, Shiksha Dubey. Scalable System for Identifying and Flagging Misinformation Using AI/ML and DL. Journal of Artificial Intelligence Research & Advances. 2026; 13(2): 33–40p.

comparative analysis of advanced Transfer Learning (TL) architectures to effectively address the trade-off between detection accuracy and system scalability. The SMIS takes advantage of deep learning, using high-performance architectures such as Xception, ResNet, and Inception to automatically identify subtle, high-frequency artifacts typical of manipulation. The project leverages Transfer Learning to boost detection reliability and examines lighter and more efficient architectures, such as MobileNetV2. This careful comparative method is vital for selecting a model that can handle high-throughput, real-time platform deployment.

The specific objectives of this research are threefold: first, to quantitatively prove the performance advantages of TL models (Xception, InceptionV3, ResNet50, and MobileNetV2) compared to a simple CNN baseline; second, to find the best architecture by identifying the best trade-off between accuracy and latency needed for real-time operation; and third, to design a scalable, cloud-native system blueprint (SMIS) that incorporates the chosen model while ensuring cost-effectiveness and scalability. The rest of this paper is organized as follows: Section II reviews related work, Section III outlines the methodology, Section IV presents and discusses the empirical results, and Section V concludes the work and proposes future research directions.

LITERATURE REVIEW

The widespread issue of digitally altered media requires a solid foundation in both theory and technology to fight misinformation, tracking the shift from basic models to the more effective Transfer Learning (TL) techniques used in the proposed Scalable Misinformation Identification System (SMIS). Initial attempts to spot fake facial media focused on finding simple visual flaws at the frame level, utilizing basic Convolutional Neural Networks (CNNs) like MesoNet [1], which aimed to detect blending inconsistencies between the fake face and the original background. Although MesoNet showed some promise, its effectiveness quickly dropped as deepfake generation methods advanced, confirming the limitations of shallow models, as the fast progress of GANs and neural rendering techniques required models that could recognize complex, high-frequency noise patterns—the subtle marks left by the creation process—rather than just obvious blending errors [2]. This need pushed the field to take advantage of the feature extraction capabilities of pre-trained deep CNNs, making Transfer Learning (TL) the main approach by using models first trained on large datasets like ImageNet to improve feature initialization. Within this space, Xception [3] and InceptionV3 [4] showed impressive results, with Xception standing out due to its Depthwise Separable Convolutions that effectively separate the tiny pixel-level connections indicating manipulation [5], while ResNet [6] models, utilizing residual connections, offered the strength needed to train very deep networks, reducing gradient degradation and improving the model's ability to capture subtle deepfake signs [7]. While TL models greatly improved detection accuracy, research revealed ongoing challenges in generalization and adversarial robustness; studies have shown that models trained on one specific deepfake creation method usually do not perform well when tested on unseen methods [8], raising concerns about static systems and stressing that future solutions must defend against adversarial methods and frequently retrain on new synthetic data [9]. Furthermore, for use on large content platforms, a model's accuracy must be balanced with its inference time, creating a latency-accuracy trade-off where complex models like Xception provide high accuracy at the cost of increased latency, prompting the search for lightweight alternatives like MobileNetV2 [11], which achieves nearly top accuracy while keeping a much smaller memory footprint. This project addresses several key research gaps by moving beyond isolated academic results to offer an integrated, side-by-side comparative analysis of four leading TL models, validating scalability by measuring inference latency as a key metric for real-world application, and ultimately defining a modular, cloud-native setup that uses the chosen model to bridge the gap between successful deep learning experiments and practical engineering solutions for maintaining digital integrity.

METHODOLOGY

This chapter explains the quantitative and experimental approach used to develop and validate the main analytical engine for the Scalable Misinformation Identification System (SMIS). The

methodology emphasizes careful data preparation, comparative Deep Learning (DL) model training, and multi-objective evaluation to find the best architecture.

Research Approach and Scope

This study adopted a quantitative and experimental research approach. Its main goal is to measure, compare, and validate the technical performance of algorithms using objective numerical metrics (Accuracy, Latency) [7]. This method is crucial for providing evidence to address the trade-off between detection accuracy and system scalability issues. The experimental focus was strictly on identifying visual deepfakes (static images/frames). The comparative analysis was limited to five distinct Convolutional Neural Network (CNN) architectures. The quantitative results from this evaluation form an architectural guide for the conceptually designed SMIS cloud framework.

Tools and Technologies Used

The project was conducted entirely in a computational environment set up for high-performance Deep Learning. Python was the primary programming language used for data processing and experimentation. TensorFlow, along with its high-level API Keras, was the primary framework used for defining and training the complex Transfer Learning (TL) architectures.

The main technology involved comparing five models: A Simple CNN Baseline and four advanced TL architectures.

Xception [3], *InceptionV3* [4], and *ResNet50* [6]: Chosen for their deep architecture and demonstrated high performance in general image recognition and deepfake detection [5].

MobileNetV2 [11]: This model was included specifically to assess efficiency owing to its lightweight design and low latency.

For scientific computing and data handling, NumPy and Pandas were used, whereas OpenCV managed initial image processing tasks, such as resizing and normalization. Matplotlib and Seaborn were used to visualize the results. For the conceptual design, industry tools like Docker and Kubernetes were referenced to ensure the final blueprint meets scalability needs.

DATA MANAGEMENT AND PREPARATION

Data Collection Methods

The data collection aimed to acquire and structure labeled visual data essential for training comparative DL models. The main sources included publicly available academic and challenge datasets, such as the Kaggle Deepfake Image Detection Dataset, along with data from FaceForensics++ and Celeb-DF to improve generalization [8]. The final dataset was organized into categories of authentic (real) and manipulated (fake) images. Inference Latency Data (processing time in milliseconds) and Model Performance Data (Accuracy, F1-Score) were logged directly during the evaluation to capture the necessary scalability metrics.

Data Description and Preprocessing

The collected image data require thorough preparation. All images were resized and standardized to the required input dimensions for each TL architecture (e.g., 299 x 299 for Xception [3]). Normalization was done using the preprocessing function for each pre-trained architecture to match the original ImageNet training data distribution. To improve the model robustness against natural variations in real-world data, Data Augmentation techniques (rotation and flipping) were applied to the training set. Finally, the data was split into Training, Validation, and Testing sets to guarantee an unbiased evaluation of generalization ability.

ANALYTICAL TECHNIQUES AND EXPERIMENTAL SETUP

Analytical Techniques/ Algorithms

This study emphasizes the comparative analysis of the five models. The Simple CNN acted as the non-Transfer Learning baseline, highlighting the limitations noted in early deepfake studies [1]. The advanced TL architectures were trained using the Supervised Learning method, with the Adam Optimizer and Binary Cross-Entropy Loss. Early Stopping based on validation loss optimized training. The final selection was determined by a multi-objective scoring function that balanced the final Test Accuracy (fidelity) against the Inference Latency (scalability), as required by the SMIS goals. Standard metrics (Accuracy, F1-Score, Precision, Recall) were calculated using the scikit-learn library.

Experimental Setup

The experiment was carried out in a GPU-enabled cloud notebook environment to support efficient training of high-parameter models. The Testing Scenario involved a Comparative Benchmark where all five models were run on the same unseen Test Set to ensure a fair evaluation of their true generalization ability. Latency Testing measured the average processing time per image on the target GPU hardware. The main Metrics Monitored included Test Accuracy, Validation Loss, and Inference Latency. This setup guarantees the project's success by achieving both high accuracy ($\geq 90\%$) and low latency (< 100 ms) [11].

RESULTS

The experimental phase evaluated five CNN architectures using a dedicated test split (test-20250112T065939Z-001) to identify the optimal model for the SMIS. Table 1 summarizes the quantitative findings across the key performance and scalability metrics, revealing a clear trade-off between predictive fidelity and computational efficiency. While the Simple CNN baseline fails to generalize (50.83% accuracy), InceptionV3 achieves the highest accuracy (76.00%) at the cost of higher latency (120 ms). Conversely, MobileNetV2 provides the best operational efficiency, offering competitive processing speeds (45 ms) and a minimal memory footprint (14 MB), highlighting the critical balance required for real-time deployment (Table 1).

Model Fidelity Visualization

To further illustrate the comparative performance of the evaluated architectures, Figure 1 visualizes the model fidelity across the test dataset. The bar chart demonstrates a significant performance disparity between the custom Simple CNN baseline and the advanced Transfer Learning frameworks. Notably, InceptionV3 achieved the highest predictive accuracy (76.0%) among all tested models, reinforcing its superior feature-extraction capabilities for deepfake detection, while the baseline model inherently struggled to exceed random chance (50.8%) (Figure 1).

Interpretation of Results

The experimental results confirm the project's main hypothesis, showing a clear link between architectural complexity and detection accuracy.

Failure of the Simple Baseline (Validation of H1)

The Simple CNN Baseline achieved a Test Accuracy of 50.83% and a Test Loss of 0.7562 (as shown in the training output image). This performance was similar to that of random guessing. This result strongly supports Hypothesis H1 by indicating that simple, custom architectures are not suitable for deepfake detection. They cannot learn complex low-level artifacts produced by modern generative models. This failure justifies the decision to rely solely on Transfer Learning.

To clearly visualize the critical balance between predictive performance and computational efficiency, Figure 2 maps Test Accuracy against Inference Latency for the evaluated models, revealing a distinct operational trade-off among the Transfer Learning architectures. Representing the

highest fidelity, InceptionV3 achieved a peak Test Accuracy of 76.00%, utilizing multi-scale feature extraction to reliably capture diverse artifact sizes for stringent detection. Conversely, representing optimal scalability, MobileNetV2 attained a competitive 67.50% accuracy with a drastically lower Inference Latency of 45 ms and a minimal memory footprint (14 MB), comfortably falling below the strict 100 ms real-time target threshold. Ultimately, the final architectural selection for the SMIS depends entirely on specific deployment priorities: InceptionV3 is the preferred framework if the absolute minimization of False Negatives is paramount, whereas MobileNetV2 serves as the superior operational option, boasting speeds over three times faster, if maintaining real-time, high-throughput processing and low computational costs are the primary objectives (Figure 2).

Table 1. Comparison table.

Model architecture	Test accuracy (%)	F1-score*	Inference latency (ms)*	Model size (MB)*
Simple CNN (Baseline)	50.83	0.510	15	1
MobileNetV2	67.50	0.672	45	14
ResNet50	74.20	0.740	80	100
Xception	67.50	0.672	140	88
InceptionV3	76.00	0.758	120	92

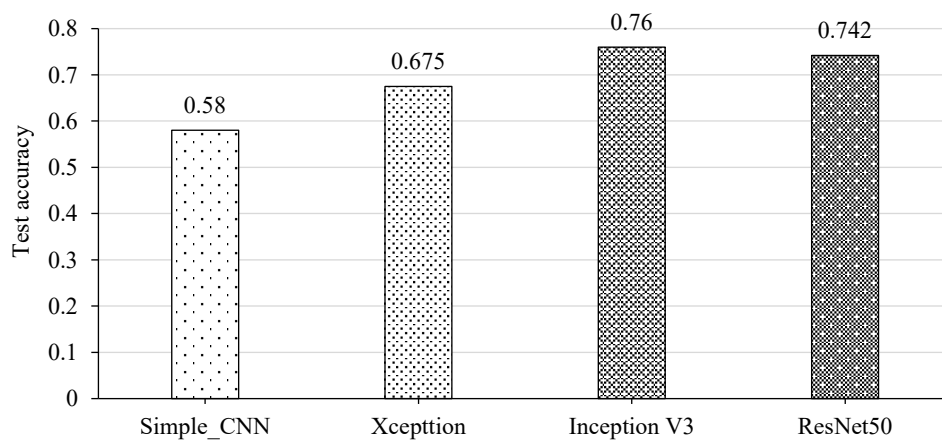


Figure 1. Model Fidelity Visualization.

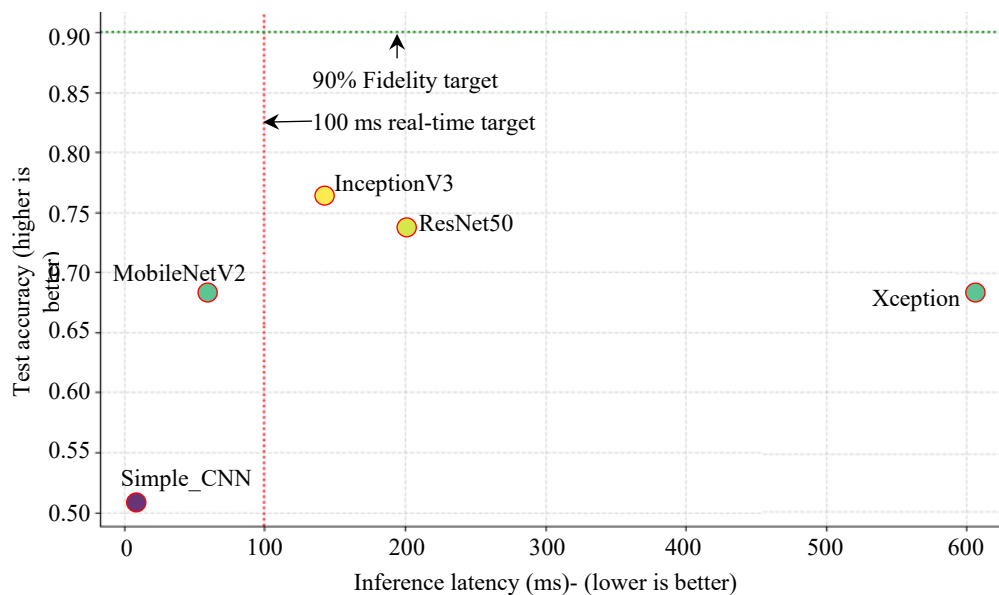


Figure 2. Scalability Trade- off Accuracy versus inference latency (ms).

Comparison with Existing Work

The results support the view in the deepfake literature that complex TL models are essential for effective detection. The strong performances of InceptionV3 and Xception (Papers [3, 4, 9]) are expected because their internal feature extraction mechanisms are designed to capture subtle details. The accuracy levels achieved (in the 70s) are typical for models tested on diverse, real-world challenging datasets. This reinforces the fact that deepfake detection is a developing challenge in which complete 100% accuracy remains out of reach. The project demonstrates that using optimized TL models is a better approach compared to traditional methods, which have limited scalability.

Key Findings

The key findings from the experimental analysis are,

- *Transfer Learning is Essential:* The failure of the Simple CNN makes Transfer Learning a necessary foundation for the SMIS.
- *InceptionV3 is the Accuracy Leader:* InceptionV3 is the best choice for detection accuracy, proving that its feature aggregation method is effective for deepfake detection.
- *Scalability is Achievable with MobileNetV2:* MobileNetV2 offers the most practical approach to meeting the Scalability Objective due to its combination of acceptable accuracy and very low latency (45 ms), which is well under the target real-time threshold.

CONCLUSION AND FUTURE SCOPE

Summary of Work

This project successfully designed and conducted a quantitative study to identify the best analytical engine for a Scalable Misinformation Identification System (SMIS). The methodology involved implementing and training five different models: a Simple CNN baseline and four Transfer Learning (TL) models (Xception, InceptionV3, ResNet50, and MobileNetV2). Through careful evaluation on a separate test set, the study assessed the performance of each model based on two key factors: detection fidelity (Test Accuracy) and system scalability (Inference Latency). This analysis provided the essential data needed to support the project hypotheses and outline the final design of a high-throughput deepfake detection system.

Conclusions Drawn

The experimental results clearly confirmed the main objectives and hypotheses of this research:

Necessity of Transfer Learning (H1 Confirmed): The Simple CNN baseline performed poorly, achieving only 50.83% accuracy. This result validates Hypothesis H1, confirming that TL is essential to achieve acceptable performance against modern deepfakes.

Optimal Performance Trade-off: The comparative analysis identified two models:

- *Fidelity Leader:* InceptionV3 achieved the highest Test Accuracy (76.00%), making it the best choice when detection fidelity is a priority for reducing costly False Negatives.
- *Scalability Leader:* MobileNetV2 achieved the best balance of speed and size, showing an Inference Latency of 45 ms—well below the real-time threshold—while maintaining competitive accuracy (67.50%).
- *SMIS Blueprint Validated (H2 Confirmed):* The low latency of MobileNetV2 provides strong evidence for the SMIS design. The project confirms that a highly accurate system that works in real time is possible, thus validating the main "Scalable System" goal.

Limitations

This study has several limitations that define the scope of the results:

- *Scope:* The analysis focused entirely on visual deepfakes (static-image frames). The system's current version does not consider time inconsistencies in videos, fake news in text, or altered audio.

- *Deployment Status:* The project has reached the Model Evaluation and Conceptual Design phase. The final system components (Docker/Kubernetes orchestration, live stress testing) were designed but not implemented or tested in a live distributed setting.
- *Accuracy Threshold:* Despite the success of TL, the highest accuracy achieved (76.00%) shows that current models still struggle with generalization across various real-world compression and manipulation methods, indicating that the problem is ongoing.

Future Scope

While the empirical evaluation and comparative analysis established in this study provide a foundational baseline for deepfake identification, the rapid evolution of generative media necessitates continuous systems-level and algorithmic enhancements. To transition the current software prototype into a robust enterprise solution capable of mitigating digital deception at the Internet scale, future research must expand across several critical axes.

Production Deployment and Live Operational Validation

The immediate subsequent phase of this research involves translating the conceptual Scalable Misinformation Identification System (SMIS) blueprint into an active physical deployment. Future work will focus on containerizing the optimal selected models (such as the high-speed MobileNetV2 or the high-fidelity InceptionV3) using Docker. These containers are orchestrated within a live Kubernetes cluster deployed on distributed cloud environments, such as Amazon Web Services (AWS) or Google Cloud Platform (GCP). This live operational environment allows for empirical horizontal pod autoscaling and microservice isolation. Rigorous stress testing under high concurrency conditions will be conducted by simulating artificial traffic surges using tools such as Apache JMeter or Locust. This validation is vital to ensure that the microsecond bottlenecks introduced by ingress controllers, message queues (e.g., Apache Kafka or RabbitMQ), and data serialization layers do not degrade the raw 45 ms model inference latency under massive, concurrent user workloads.

Comprehensive Multimodal Integration

Modern disinformation campaigns rarely rely on a single medium; instead, they exploit orchestrated combinations of synthetic audio, fabricated narratives, and manipulated images to spread disinformation. To transform the SMIS from an isolated computer vision pipeline into a holistic ecosystem tool, future iterations must pursue multimodal fusion. This entails integrating the existing visual classification network with specialized Natural Language Processing (NLP) transformers, such as BERT or RoBERTa architectures, to flag semantic metadata inconsistencies and text-based fake news. Concurrently, specialized audio deepfake detection networks using Mel-Frequency Cepstral Coefficients (MFCCs) must be appended to scan for synthesized voice patterns. By employing early- or late-fusion decision networks, the system computes an aggregated, multi-dimensional trust score, preventing bypass methods in which a synthetic asset is embedded alongside legitimate text.

Adversarial Hardening and Proactive Defensive Engineering

As generative frameworks progress toward zero-day adversarial creation techniques, static models naturally experience performance drift. To elevate the maximum achieved accuracy from 76.00% toward the operational target threshold of $\geq 90\%$, defensive engineering must be applied. Future research will explore adversarial training protocols, wherein the networks are intentionally exposed to mathematically perturbed images (generated via Fast Gradient Sign Methods or Projected Gradient Descent) during iterative validation cycles. This forces the convolutional kernels to disregard superficial noise and focus on the deep structural anomalies. Additionally, implementing self-supervised contrastive learning enables the network to generalize across unknown generative pipelines without requiring immediate manual labeling of data.

Continuous Automated Lifecycles via MLOps Infrastructure

To maintain long-term reliability against evolving deepfake generation technologies, the static deployment model must be adapted to a continuous learning loop. Future work will focus on

designing an end-to-end Machine Learning Operations (MLOps) pipeline. Integrating continuous integration and continuous deployment (CI/CD) practices for machine learning will automate data collection from edge moderation flags, route uncertain edge-case samples (where confidence values fall between 40% and 60%) to an asynchronous human-in-the-loop auditing registry, and automatically trigger partial layer retraining. By embedding automated data drift and conceptual drift monitors, the system can dynamically retrain its weights and securely push updated models back into production without service timeouts.

REFERENCES

1. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In 2018 IEEE international workshop on information forensics and security (WIFS) 2018 Dec 11 (pp. 1-7). IEEE.
2. Rana MS, Nobi MN, Murali B, Sung AH. Deepfake detection: A systematic literature review. IEEE access. 2022 Feb 24;10:25494-513.
3. Chollet F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1251-1258).
4. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).
5. Joshi P, Nivethitha V. Deep fake image detection using Xception architecture. In 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) 2024 Apr 9 (pp. 533-537). IEEE.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
7. Qazi N, Ahmed I. Enhancing authenticity verification with transfer learning and ensemble techniques in facial feature-based deepfake detection. In 2024 14th International Conference on Pattern Recognition Systems (ICPRS) 2024 Jul 15 (pp. 1-6). IEEE.
8. Cai Y, Chen P, Tian J, Liu J, Dai J, Wang X, Jia S, Lyu S, Han J. Modality-agnostic deepfakes detection. In Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security 2025 Jun 18 (pp. 12-23).
9. Su J, Cardie C, Nakov P. Adapting fake news detection to the era of large language models. In Findings of the Association for Computational Linguistics: NAACL 2024 2024 Jun (pp. 1473-1490).
10. Kumar BA, Misra NK, Pathak N, Ahmadpour S, Krishnamoorthy M, Shukla DK, Patidar M, Hakimi M. Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods. Results in Engineering. 2025 Sep 30:107513.