# DEEPFAKE IMAGE DETECTION USING MACHINE LEARNING

A.Madhav Reddy-221FA04260
*Department of Computer Science And Engineering*
*Vignan's Foundation for Science, Technology and Research*
Guntur, India

P.B.S.V.Karthikeya-221FA04342
*Department of Computer Science And Engineering*
*Vignan's Foundation for Science, Technology and Research*
Guntur, India

M.Deepika-221FA04594
*Department of Computer Science And Engineering*
*Vignan's Foundation for Science, Technology and Research*
Guntur, India

M.Abhinaya-221FA04678
*Department of Computer Science And Engineering*
*Vignan's Foundation for Science, Technology and Research*
Guntur, India

## I. INTRODUCTION

The spread of deepfake images, made possible by the popularity of social media, poses a major threat to the authenticity of online information. The manipulated images are part of the spread of misinformation, which undermines public confidence in digital content. In response to this emerging issue, this project seeks to create an advanced machine learning-based system for automatic detection and labeling of deepfake images. By examining minute, usually unnoticeable details, the system will correctly distinguish between genuine and forged photographs. Having been trained on large datasets of genuine and synthetic images, it will learn to recognize varied types of forgery with increased reliability. Additionally, the support for real-time detection by the system will enable users to make rapid determinations of image authenticity, preventing the dissemination of deceptive content and strengthening digital security.

**Background and Significance**

Growing sophistication of deepfake technology requires sophisticated detection techniques. Machine learning represents a powerful method of analyzing sophisticated image patterns and detecting nuanced inconsistencies that betray manipulation. Automatic and accurate detection of deepfakes is key to sustaining the integrity of digital media and preventing the exploitation of manipulated content for ill use.

**Machine Learning in Digital Media**

Machine learning models are increasingly being utilized for the different aspects of online media, such as video and image analysis, content moderation, and detecting fraud. Their capacity to learn from extensive datasets and evolve based on changing patterns makes them ideally suited for addressing the dynamic nature of deepfake detection.

**Objectives and Scope**

The main aim of this project is to create a strong machine learning-based system able to effectively identify deepfake images. The scope consists of:

1.Creating a model that can identify real and fake images by examining image features.
2.Enabling real-time detection features.
3.Evaluating the model's performance on diverse datasets.
4.Investigating methods to enhance the robustness of the model against advancing deepfake technologies.

**Challenges in fake image detection**

1.The quick evolution of deepfake generation methods.
2.The subtle character of certain manipulations, which may be hard to spot.
3.The requirement of real-time processing to manage massive amounts of content online.
4.Variability of image compression and resolutions available online.
5.The capability of generalization to deepfakes developed using undisclosed processes.

**Applications of ML for deepfake image detection** Machine learning can be used in a range of applications, such as:

1.Social media automated fact-checking and content moderation. Legal investigations and digital forensics.
2.Media verification and provenance tracking.
3.Browser extension and tool development for real-time deepfake detection.
4.Security solutions to stop deepfakes from being utilized to circumvent biometric security.

## II. LITERATURE REVIEW

Andreas talked about the fact of modern image manipulations and that it is hard to identify them automatically or manually. Once the data have been acquired, it is manipulated, and CNNs are used to establish if the image is actual or not.[1]

Yuezun Li emphasized large-scale dataset requirements for testing and developing deepfake detection models. Deepfakes that already exist are not visually as high in quality as what can be downloaded on the internet. This has been answered with Celeb-DF3, a difficult, large-scale deepfake video dataset for the detection of deepfakes.[2]

Brian introduced the DFDC dataset, the largest publicly released face swap video dataset, comprising more than 100,000 videos of 3,426 paid actors. The dataset is created with various deepfake and GAN-based approaches, thus making it suitable for evaluating detection models.[3]

Ricard proposed a machine learning-based method that identifies forged videos from the FaceForensics++ dataset with 90 percent accuracy. The method relies on frequency analysis, which detects different behaviors at high frequencies, making it effective in recognizing artificial image content.[4]

Ruben gave an overview of methods to identify and manipulate face images, such as four categories of face editing: whole face modification, identity swapping, feature modification, and expression swapping. The survey emphasizes the need to examine different forms of facial manipulation for deepfake detection.[5]

Nicol'o used CNN techniques to detect face changes in more than 10,000 videos and proved the strength of CNNs in detecting fake media. .The research refers to facial manipulation methods in existence and how they affect detection.[6]

Wanying Ge proposed the application of SHapley Additive exPlanations (SHAP) to explain machine learning model predictions. SHAP is the feature contribution to the prediction of a model, which makes the deepfake detection system more explainable.[7]

Chunlei Peng suggested giving different fidelity scores to real and forged face data, enhancing the model's capability to detect complicated samples. The approach takes perceptual forgery fidelity into account and gives discrete values to facial data according to various attributes.[8]

Tianchen showed that unique source characteristics retained and recovered throughout deepfake generation algorithms can be utilized for detection. The approach estimates self-consistency by examining local source characteristics at multiple points in the image.[9]

Bojia presented the WildDeepfake dataset, with 7,314 face sequences from 707 deepfake videos collected from the web. Attention-based Deepfake Detection Networks (ADDNets) were suggested to enhance detection performance on real-world deepfakes.[10]

Kaede introduced self-blended images (SBIs) as artificial training data to simulate forging artifacts. SBIs are produced by mixing slightly altered source and target images from real images, which enables the model to learn forgery patterns.[11]

Shichao proposed the FST-Matching Deepfake Detection Model to enhance forgery detection on compressed videos. The model improves artifact feature learning under binary supervision, leading to better detection.[12]]

Anubhav Jain put forth a system to detect deepfakes utilizing synthetic data obtained with StyleGAN3. The model minimizes bias and increases interpretability without needing real data.[13]

Fatima trained a data set of 9,000 images for 150 epochs, observing that ResNet50 performed 100 percent training accuracy, 99.18 percent validation accuracy, and 99 percent test accuracy and hence was the best architecture.[14]

## MOTIVATION FOR DEEPFAKE IMAGE DETECTION

The driving motivation for this study arises from the pressing requirement of safeguarding digital information integrity and addressing the ever-increasing risk of deepfakes. With such fake images becoming more and more convincing as well as omnipresent, there is an urgent need to secure societal confidence as well as authenticity of digital content. Below are some main drivers for this work:

Preservation of Informational Accuracy and Confidence: Through building stable automated deepfake detection approaches, we are able to prevent deception and reclaim users' faith in online media. This is fundamental in securing an open and fair information space, and for mitigating loss of public confidence in digital information.

Enhanced Effectiveness and Scalability of Detection These automated processes could potentially speed up digital image processing significantly, allowing for rapid detection of deepfakes in large bodies of online content. This scalability is critical in attempting to keep pace with the exponential growth of manipulated media and be able to combat it effectively.

Enhancing Digital Security and Accessibility This research paves the way for the creation of affordable and reliable deepfake verification software, which enables individuals and businesses to verify digital content. This not only enhances digital security but also literacy, especially where misinformation could result in serious consequences.

Facilitating Innovative Applications in Media Authenticity: The integration of machine learning and computer vision in deepfake detection provides promising avenues for media forensics, content origination tracking, and real-time authentication. These technologies can assist in developing efficient digital protection mechanisms and enhancing the transparency of online communication.

## III. METHODOLOGY

The approach suggested here attempts to detect deepfakes using machine learning technologies with a specific focus on Convolutional Neural Networks (CNN). An efficient machine learning model that can distinguish between real and manipulated images is attempted to be created. Image pattern analysis, inconsistencies in facial structures, and artifacts formed as a result of deepfake generating processes are the approaches it adopts. For achieving improved accuracy and performance, comparisons with current machine learning models. Machine learning, in the form of CNNs, is used for image classification and deepfake detection. A CNN consists of various layers such as convolutional, activation (ReLU), pooling, and fully connected layers. These layers enable feature extraction and classification, hence enabling CNNs to perform deepfake detection tasks. The convolutional layer identifies major features of an image such as edges and texture with the help of filters (kernels). The ReLU layer enables non-linearity for feature learning enhancement. The pooling layer avoids overfitting and minimizes dimensionality by down-sampling. The fully connected layer pools the extracted features to give the final classification output. There are three convolutional layers with dropout, max pooling, and fully connected layers in the model. The layers assist the network in learning useful patterns and outliers from deepfake images. The proposed model classifies an image as real or deepfake based on a CNN architecture designed for maximum accuracy. The initial convolutional layer employs 32 filters of size 3x3. The second and third convolutional layers employ 64 and 128 filters, respectively. Max pooling is employed after each convolutional layer with pool size 2x2. 0.3 dropout is employed to avoid overfitting and improve generalization. A dense layer is employed after flattening the feature maps to produce the final classification output. The Adam optimizer is used for model training with binary cross-entropy as loss function, which is suitable for binary real vs. fake image classification. Deepfake images typically possess soft artifacts and inconsistencies which may be targeted particularly for their detection. Feature extraction techniques are utilized within the model to enhance classification efficacy. Frequency analysis is applied to high frequency noise and fake texture prevalent in deepfakes. Facial landmark analysis inspects large facial eyes, nose, and mouth for unnatural distortions. Edge detection mechanisms identify blurred or deformed edges and color discontinuities are assessed to identify unnatural light and shadow artifacts. In order to make the model strong, the proposed CNN model is compared with other machine learning models, such as VGG16, a deep CNN image classification model with high accuracy; ResNet50, a residual network structure that tries to eliminate vanishing gradient issues; and EfficientNet, a computationally efficient and scalable CNN model optimized for computational complexity. All the models are trained on real and deepfake image datasets and their performance tested on the basis of accuracy, precision, recall, and F1-score. The approach in this paper strives to make the deepfake detection system more accurate, improve cybersecurity policies by avoiding misinformation, aid forensic practitioners to identify manipulated media, and maximize the detection mechanisms through continuous learning and dataset revision. By utilizing CNN-based deep learning algorithms and feature analysis methods, the current approach ensures the development of a successful and accurate deepfake image detection system.

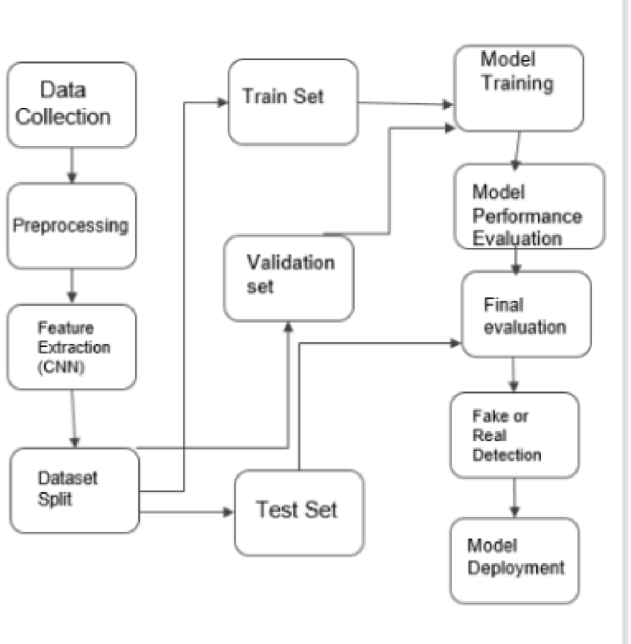


Fig. 1. Block Diagram for DeepFake Image Detection

## IV. PROPOSED SYSTEM

Deepfake image classification is an important function in digital media, and to achieve this with precision, machine learning techniques are required to properly identify real from artificial images. There are key steps involved from data collection through model assessment as detailed below:

**1. Input Data and Features:**

Data Harvesting and Organization: The dataset is highly selectively compiled and sorted into three individual folders i.e.”real” with true human face images, ”fake” containing deepfake images, and ”test-images” for evaluating the performance of the trained model. Dataset Balancing: For preventing the biasing of training, it remains balanced with an equal quantity of real and fake images. Feature Extraction Using HOG: Feature extraction uses the Histogram of Oriented Gradients (HOG) technique. The technique converts the images into grayscale and extracts texture features retaining the information pertaining to local object appearance and form.HOG parameters are given 9 orientations with 8x8 pixels per cell and 2x2 cells per block, thus yielding maximum utility for the detection of small deepfake artifacts.

**2. Data Preprocessing:**

Image Resizing: Each of the images is resized into 128 * 128 pixel images to keep input size to the model constant at all times.

Standard Color Space: Images are saved within the RGB standard color space in order to be color space-compatible.

NumPy Array Conversion: Images are converted to NumPy arrays that perform numerical operations and facilitate easy processing.

**3. Construction of the Model:**

SVM Classifier Choice: Trained with linear kernel Support

Vector Machine (SVM) classifier.

SVMs are best suited for binary classification problems as well as for high-dimensional space, and are therefore the optimal tool to use for this problem scenario.

Model Parameter Choices: SVM model with probability estimation, and standard regularization.

**4. System Methodology:**

Data Loading and Feature Extraction: The process starts by loading images from the dataset and extracting HOG features. Data Splitting: The data is split into 80% training and 20% test using the train-test-split function, making the model evaluation strong. Training method: The SVM classifier is trained on the training data so that it learns to predict the image classes based on texture patterns and local gradients.

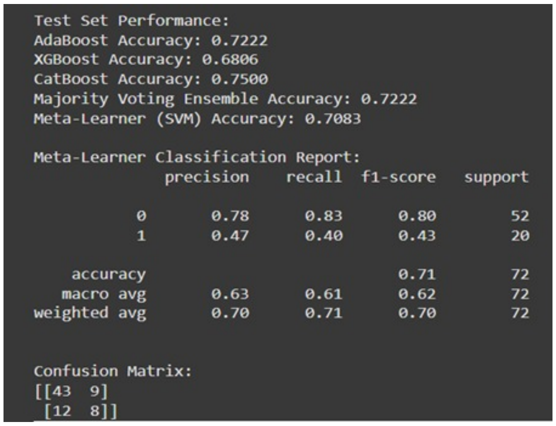
Testing Method: The trained model is tested with the test set, and performance measures such as accuracy score, precision, recall, and F1 score are computed. New Images Prediction: The system does the preprocessing and the feature extraction, and the model gives a prediction (real or fake) together with a score of confidence.

Result Visualization: The results of the prediction are visualized with matplotlib, showing the image with the percentage of confidence overlaid.

**5. Model Evaluation:**

Performance Metrics: The model has been validated with accuracy score, precision, recall, and F1-score which gives a general overview about its efficiency.

Classification Report: A classification report shall be submitted for extended status in the sense of the model's performance class-wise (real and fake) wise.



```
Test Set Performance:
AdaBoost Accuracy: 0.7222
XGBoost Accuracy: 0.6806
CatBoost Accuracy: 0.7500
Majority Voting Ensemble Accuracy: 0.7222
Meta-Learner (SVM) Accuracy: 0.7083

Meta-Learner Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.83      0.80        52
           1       0.47      0.40      0.43        20

    accuracy                           0.71        72
   macro avg       0.63      0.61      0.62        72
weighted avg       0.70      0.71      0.70        72


Confusion Matrix:
[[43  9]
 [12  8]]
```

Fig. 2. Performence Outcomes

**6. Resulted Confusion Matrix**

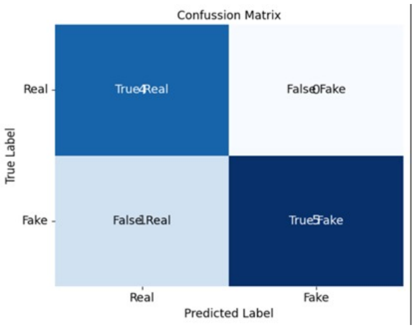


Fig. 3. Confusion Matrix

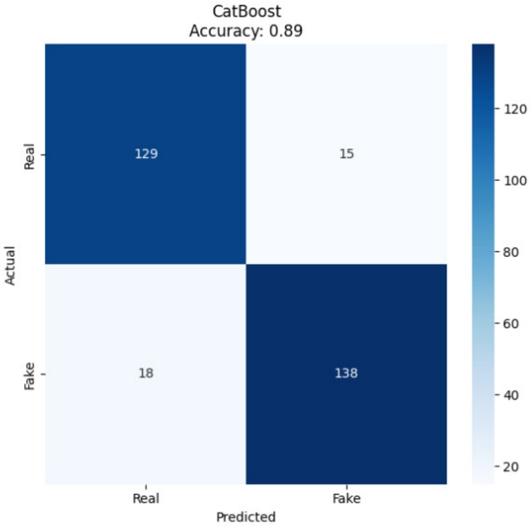


Fig. 4. CatBoost Confusion Matrix

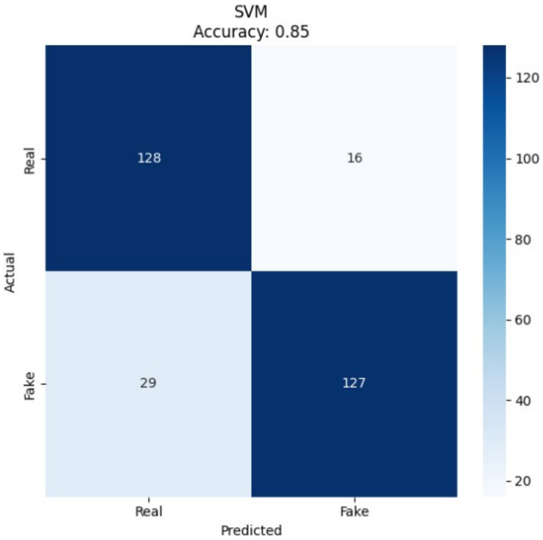


Fig. 5. SVM Confusion Matrix

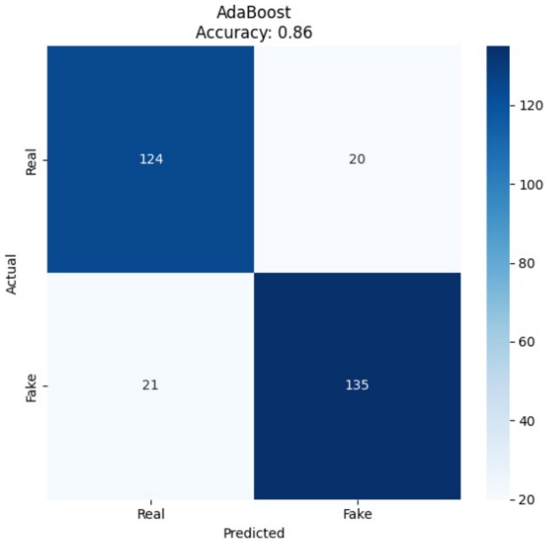


Fig. 6. AdaBoost Confusion Matrix

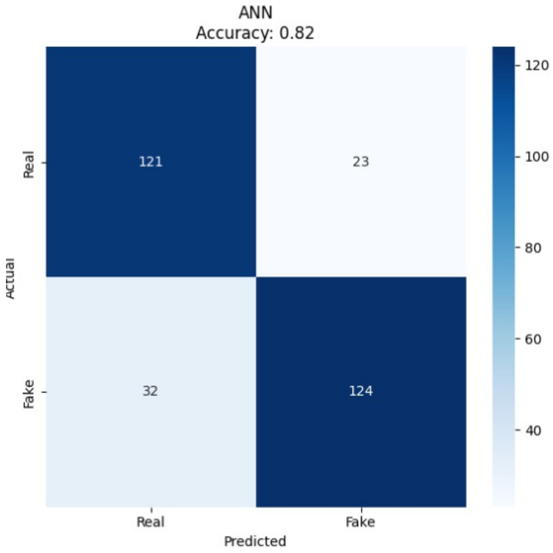


Fig. 7. ANN Confusion Matrix

### 7. Testing Vs Training

Deepfake image detection is a highly critical task that requires suitable machine learning models to separate real and fake images. In this work, we tried AdaBoost, XGBoost, CatBoost, SVM, and ANN models.

XGBoost and CatBoost achieved ideal training accuracy (1.00) but with slight overfitting, while AdaBoost and SVM had a good balance between test and training accuracy.

ANN achieved the low test accuracy of 0.82, which shows the need for improvement. Performance can be improved through regularization, data augmentation, and hybrid modeling. These findings are helpful in choosing the best performing model for detecting deepfakes.

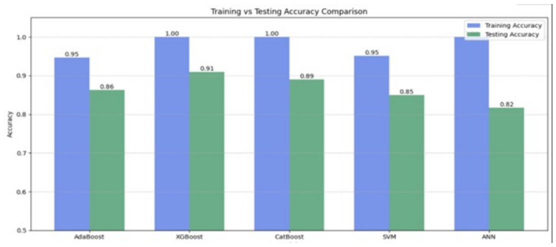


Fig. 8. Dialogical Comparison of Training vs Testing.

### 8. Constraints:

Computational Resources: The operation of the system would, to some extent, be limited by the accessible computational resources, particularly in the case of large data volumes.

Generalization toward Novel Deepfakes: The generalization capability for the introduced deepfake techniques never seen in the training set could be poor.

HOG Feature Limitations: The HOG feature method may not pick up all the tiny artifacts that are present in highly sophisticated deepfakes.

### 9. Cost and Sustainability Impact:

Computational Cost: Model training cost and run has to be considered, especially in real-time settings.Data Storage: Storage requirements for the dataset and trained model will have to be evaluated against sustainability.

Scalability: System scalability to support larger data sets and more volume of real-time data needs to be considered.

Maintenance and Updates: The long-term maintenance and updates of the system, such as retraining with new data and keeping pace with evolving deepfake techniques, should be considered in the sustainability and cost assessment.

### 10. Future Directions:

CNN Integration: Defining the integration of Convolutional Neural Networks (CNNs) to improve performance on more complex datasets and to improve feature extraction.

Advanced Feature Engineering: Investigating advanced feature engineering techniques to identify more subtle deepfake artifacts.

Optimization for Real-time Processing: Optimizing the system to enable efficient real-time processing deepfake detection in evolving conditions.

## V. CONCLUSION

For this project, deep image forgery detection has been performed via artificial intelligence-based systems and various machine learning approaches in feature extraction mechanisms that can detect any manipulated image. Some models tested include SVM, XGBoost, and ANN. Among the results, the XGBoost based model performed optimally with a testing accuracy of 91%. The result showed the efficiency of machine learning-based classifiers for detecting deepfakes in images. The model ensures accurate detection with computational effectiveness. More future work can delve into more sophisticated deep learning architecture, incorporating feature extraction methods, and bigger data sets on improving accuracy and resilience.

## VI. REFERENCES

Deep Learning-Based Deepfake Image Detection: This paper investigated the performance of deep learning techniques for deep fake image detection, focusing on Convolutional Neural Networks and Generative Adversarial Networks. The study suggested the feasibility of such models for detecting manipulated facial images by studying spatial and temporal inconsistencies. It was proven that more recent architectures such as XceptionNet and EfficientNet are able to enhance the detection accuracy without degrading computational efficiency. This will offer effective deep fake content detection, which is of utmost importance in combating misinformation as well as digital media security.[1]

Deepfakes and Beyond: The comprehensive analysis done by the researchers was regarding facial manipulation and detection techniques, categorizing up-to-date deep fake detection approaches into pixel-based, physiological-based, and temporal-based methods. The authors highlighted that part of the reason for the improvements in detection robustness have come about due to the advances in deep learning, including RNNs and architecture-based transformations. The study also offers evidence for the urgent need for real-time deep fake detection systems.[2]

The presented paper FaceForensics++, which is one of the large datasets for deep-fake detection, for various manipulation methods such as Deepfakes, Face2Face, and NeuralTextures. It has been shown that fine-tuning pre-trained deep learning models like Xception and ResNet could greatly amplify the accuracy of detection. This also provides a benchmark for testing the methods for detection of deepfakes.[3]

The study proposed a lightweight deep learning model-MesoNet-for detecting manipulated facial videos. Unlike traditional CNNs, MesoNet focuses on mesoscopic features that capture the subtle artifacts introduced during the generation of deepfake material. The model attained high accuracy and reasonable efficiency, thus making it suitable for real-time applications.[4]

Deepfake Detection Challenge Dataset Researchers at Facebook AI developed a large-scale benchmark for training and evaluating deepfake detection models, the Deepfake Detection Challenge (DFDC) dataset. The dataset consists of deepfake videos of different genres produced by various face-swapping and reenactment techniques. Insights into the performance improvement through ensemble learning are also given. [5]

The introduction of XceptionNet is the main contribution of the Xception paper. The architecture employs depthwise separable convolutions to enhance the feature-extraction efficiency. The model clobbered traditional CNNs in deepfake detection by capturing fine-grain artifacts in manipulated images. The result demonstrates that lightweight architectures can achieve a high degree of accuracy coupled with low computational costs. [6]

Deepfake Video Detection using Biological Signals Explored by the researchers were biological signals like heartbeat and facial micro-expressions for Deepfake detection. The research states that physiological variations are often missing in deepfake videos, which can be aided to improve accuracies. Joint application of computer vision along with physiological signal analysis was successful in identifying synthetic media. [7]

Two-Stream Neural Networks for Tampered Face Detection research proposed a two-stream neural network architecture for the detection of tampered facial images. One stream processed texture-based artifacts and the other processed motion inconsistencies in manipulated videos. Fusion of both streams greatly improved the detection. [8]

Media Forensics and Deepfakes research summarizes media forensics methods for deepfake detection, comparing handcrafted features to deep learning models and hybrid approaches. The importance of diversity in datasets and adversarial robustness for real-world deepfake detection systems was highlighted by the study. [9]

Large-Scale Video Dataset for Forgery Detection in Human Faces. The FaceForensics dataset has been proposed, comprising genuine and manipulated face videos for training deepfake detection models. The paper contrasted various forgery detection methods, offering benchmarks for numerous machine learning models. The dataset is a goldmine for ongoing research in this arena. [10]