ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

A Comparative Benchmarking Study of Classical Machine Learning and Deep Learning Methods for Image-Based Deepfake Detection

Kushal Zanzari | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0009-0004-7201-6324

Jayesh Muley | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0009-0008-2164-2658

Sneha Chandravanshi | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0009-0000-8364-7157

Anshika Jain | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0009-0008-0360-6872

Taufik Hussain | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0009-0001-8332-4312

Ajay Kumar Phulre | Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, India | ORCID: 0000-0001-7457-1007

Abstract

Deepfake media provides a very genuine threat to digital trust, enabling misinformation, fraud, and impersonation of identity. This paper examines a comparison benchmark analysis between classical machine learning (ML) approaches using Support Vector Machines, Decision Tree, Random Forest, and Gradient Boosting as well as deep learning (DL) methods based on Convolutional Neural Networks (CNNs) for image-based deepfake detection. Although FaceForensics++ and Celeb-DF are video datasets, a fixed number

Received: 11.11.2025

Accepted: 24.04.2026

Published: 30.06.2026

Cite this article as:
K. Zanzari, J. Muley, S. Chandravanshi, A. Jain, T. Hussain, A.K. Phulre, "A comparative benchmarking study of classical machine learning and deep learning methods for image-based deepfake detection," ACIG, vol. 5, no. 1, 2026, doi: 10.60097/ACIG/221087

Corresponding author:
Kushal Zanzari, Department of Computing Science Engineering and Artificial Intelligence, Vellore Institute of Technology, Bhopal, Madhya Pradesh, India. E-mail: kushalzanzari2022@vitbhopal.ac.in
0009-0004-7201-6324

of face frames were extracted and processed to construct a balanced dataset of 3930 images. Experimental evaluation shows that classical ML models achieve accuracies ranging from 84% to 91%, with gradient boosting performing the best among these, while the CNN achieves the highest accuracy of 93%. Though it has shown increased precision in its capability to detect, traditional models also provide an edge in terms of computational efficiency and interpretability. The significance of this particular study has been to bring forward inherent trade-offs and provide selection guidance for deepfake models.

Keywords

support vector machine (SVM), convolutional neural networks (CNNs), computational efficiency, fake media identification

1. Introduction

Deepfake technology has increased immensely, posing an extreme threat to security, trust, and online authenticity. Deepfakes, developed based on advanced artificial intelligence (AI), can simulate visual and audio inputs to create highly realistic yet fake media. Consequently, there is more misinformation, financial fraud, and identity theft, necessitating an urgent need for effective detection measures. Traditional detection methods, including forensic analysis and rule-based systems, are incapable of handling the complexity introduced by AI-generated content. Machine learning (ML) and deep learning (DL) methods have thus emerged as possible solutions for deepfake detection. However, these methods possess certain strengths and weaknesses. Traditional ML models, like Support Vector Machines (SVM) and Decision Trees, are computationally efficient and interpretable but might lack the desired accuracy in detecting advanced deepfakes. DL architectures, like Convolutional Neural Networks (CNNs), have improved detection precision but require extensive training datasets and significant computational resources. Our study aims to provide a comparison of the performance of classical ML and DL techniques in detecting deepfakes. We test the efficiency, accuracy, and performance of the models based on benchmark datasets, such as Face Forensics++. By comparing their strengths and limitations, this study aims to find the most appropriate methodology for deepfake detection and contribute to the existing body of research in securing digital content against AI-created counterfeits. While numerous earlier studies evaluate classical versus DL-based techniques in deepfake detection, a significant amount of literature

still seems to concentrate more on accuracy metrics rather than methodological clarity, deployment issues, and decision guidance for practitioners. The current work also found that relatively less emphasis has been placed on comparisons involving reduced data scenarios, fundamental model comparisons as well as accuracy, interpretability and efficiency balances in deepfake-related studies. Accuracy, methodological clarity, reproducibility, and applicability within the current work were considered to provide a more transparent evaluation of DL-based techniques versus classical methods for deepfake detection. This study aims to compare classical ML models with a lightweight CNN for image-based deepfake detection under limited data conditions. It evaluates trade-offs in accuracy, interpretability, and computational efficiency. The hypothesis suggests that lightweight CNNs may outperform classical methods while remaining suitable for resource-constrained environments.

2. Literature Review

Deepfake technology is becoming exponentially more sophisticated, which has led to an equally significant increase in study on how to identify them. Once successful, traditional forensic techniques are not able to detect subtle alterations produced by sophisticated generative models. Consequently, the area has moved towards methods based on DL and ML. Using handcrafted features like local binary patterns (LBPs), colour texture analysis, and histogram of oriented gradients (HOG), classical ML algorithms such as SVMs, Decision Trees (DT), Random Forests (RF), and Gradient Boosting (GB) were widely used in the early stages to differentiate between real and manipulated faces based on pixel-level irregularities [1–4]. Although these models provided interpretability and computational efficiency, they were less successful in identifying high-quality, Generative Adversarial Networks (GAN)-generated deepfakes due to their dependence on shallow characteristics.

By facilitating automated feature extraction from unprocessed picture data, DL techniques – in particular, CNNs – emerged as a potent substitute to get around these restrictions. In order to detect deepfakes based on spatial artifacts, texture inconsistencies, and compression anomalies, CNNs, such as XceptionNet, MesoNet, and VGG variants, have been trained on datasets like FaceForensics++ [5–8]. On benchmark datasets, CNNs frequently achieved detection accuracy of above 90%, according to several studies that documented notable performance increases. CNN designs have also used residual connections and attention techniques to improve their focus on important face areas that are manipulable [9,10].

The advantages of both worlds may be combined using hybrid techniques, which use CNNs for deep feature extraction and powerful ML models like Random Forests and XG-Boost for classification [11,12]. These combinations maintain excellent accuracy while enabling shorter training times, improved generalisation, and more interpretable results. Certain frameworks improve the discriminative capability of tree-based classifiers without requiring a large amount of labelled data by using CNN-based embeddings as feature inputs. By combining the predictions of several base learners, ensemble approaches further improve robustness and can handle a greater range of manipulation kinds and attributes [13].

In order to detect deepfakes in video footage, recent studies have also looked at temporal information. Motion inconsistencies and irregular frame transitions are captured by methods such as 3D CNNs, Long Short-Term Memory (LSTMs), and spatiotemporal analysis, which are frequently missed by static image-based detectors [14,15]. As additional characteristics to improve detection robustness, physiological inconsistencies, such as aberrant eye blinking patterns, a lack of subtle head movements, and unnatural lip-sync, have been investigated as well [16,17].

Additionally, by using pre-trained models on large-scale datasets and refining them on deepfake-specific datasets, transfer learning has been demonstrated to enhance greatly deepfake detection. Effective cross-dataset generalisation has been shown by models like ResNet and EfficientNet, especially when paired with domain adaption strategies [18]. Attention-based CNNs and capsule networks are becoming more popular because of their capacity to preserve face part posture and spatial hierarchy, which aids in identifying minute manipulations that fool conventional CNNs [19].

The detection of audio deepfakes, which create realistic-sounding fake sounds using speech synthesis and voice cloning models, is another rapidly expanding field. In order to detect such forgeries, ML and DL techniques have been applied to waveform patterns, Mel-frequency cepstral coefficients (MFCC), and spectrogram features, enhancing visual-based detection frameworks [20].

3. Experimental Setup

3.1. Hardware

Experiments were conducted on a system with an Intel i7 CPU, 16 GB RAM, and NVIDIA GTX 1660 GPU.

3.2. Software

Python, TensorFlow, Scikit-learn, OpenCV, and NumPy were used for model development and evaluation.

4. Methodology

4.1. Dataset

This study uses FaceForensics++ and Celeb-DF datasets for image-based deepfake detection. Approximately 5–10 frames per video were uniformly sampled to reduce redundancy while preserving diversity, followed by face detection to retain frontal faces.

After preprocessing, 3930 face images were selected (1960 real and 1970 manipulated images), enabling evaluation in a small data regime. FaceForensics++ includes multiple manipulation techniques and compression levels for realistic testing.

Table 1 presents the statistical distribution of the extracted HOG features used for classical machine learning models.

Table 2 summarizes the distribution of real and manipulated samples used in this study.

4.2. Preprocessing

Prior to beginning training of a deepfake detection model, the dataset must undergo thorough preprocessing. This will include

Table 1. Feature statistics.

Feature	Count	Mean	Std	Min	Median (50%)	Max
Feature 0	3930	0.2766	0.0987	0.0000	0.3068	0.4926
Feature 1	3930	0.1857	0.1301	0.0000	0.1889	0.4478
Feature 2	3930	0.0767	0.0957	0.0000	0.0340	0.4053
Feature 3	3930	0.0327	0.0562	0.0000	0.0088	0.3803
Feature 4	3930	0.0794	0.0754	0.0000	0.0519	0.4979

Table 2. Dataset summary.

Class	Number of samples
Real (0)	1960
Fake (1)	1970

resizing, normalising, splitting, and other types of preprocessing. Usually the first action we take is to resize the images or video frames. Since a DL model will have a fixed input size, all images will be normalised to a common or standard size (i.e. 224 × 224 or 256 × 256 pixels). This lowers data volatility and boosts model data for easier learning. Since it is likely that manipulation is occurring inside the face, we manually detected and cropped the face using a face detection algorithm; so, we are only passing the model the image we wish to analyse. This will streamline the model input and help to alleviate the noise we don't want to share with the model input but focus on facial features.

After the frames have been resized, we proceed with normalisation. Normalisation enables us to normalise the values of pixels to a smaller range, such as 0 to 1 or between –1 and 1. Many DL models perform better due to a more stable and faster training process, when the input values are normalised. An additional advantage of normalisation is that it guarantees that each pixel makes a fair contribution to learning, independent from the excessive affect of elevated values. In some instances, we also subtract the mean and divide by the standard deviation of the dataset so that each feature has a similar distribution.

The data was split on the video level and split into subsets for training (70%), validation (15%), and testing (15%). Data leakage across the frame level was avoided this way to ensure that all the video frames are exclusively assigned to any subset. This is needed to ensure that our model will have the ability to generalise to previously unseen data. The training set is used to fit the model; the validation set is used to optimise and determine the best parameters; and the test is to determine the performance on completely unseen data. By ensuring that we have preprocessed the dataset in such a manner, we are assured that the deepfake detection system we are going to develop is reproducible, verifiable, and works appropriately in real world.

Figure 1 illustrates the overall image pre-processing pipeline used in this study.

4.3. Feature Extraction

Feature extraction is a very important aspect of image-based deepfake detection. This is mainly due to the fact that it is possible to differentiate between the real and manipulated images. DL-based images, such as deepfakes, usually have several

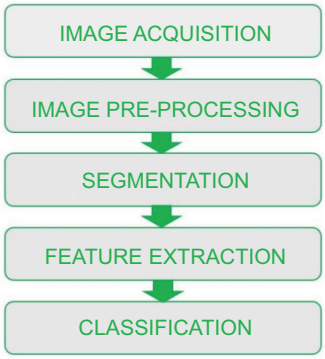

Figure 1. Image processing.

inconsistencies such as texture, inappropriate light patterns, and distortions in the faces of the involved individuals. Appropriate feature extraction is thus a necessity.

Histogram of oriented gradients feature extraction mechanism was used to extract hand-crafted feature set in this study. This mechanism successfully identifies the oriented edges in images by computing gradient directions. It helps to detect irregularities on faces by manipulating the same. All the images in this study were converted to greyscale to reduce complexity during their processing. Subsequently, they were resized to a fixed resolution of 64 × 64 pixels. Finally, the preprocessing step was followed by applying HOG feature extractors, thus allowing features to be created in a vector of a constant size. After feature vector creation, standard scaling was applied to normalise feature vectors. The aim behind feature vector normalisation was to enable proportionate contribution of all features in subsequent optimisation. The extracted features indeed provided a structured numerical representation of facial characteristics, allowing classical ML models to learn decision boundaries between authentic and manipulated images. This feature-based representation formed the basis for training and evaluating multiple traditional classifiers.

Figure 2 presents the workflow of the proposed deepfake detection framework.

1. *Handcrafted features*: These classical ML-based deepfake detection methods were fed a specific inbuilt handcrafted feature set that was extracted using an HOG descriptor. Unlike DL-based methods that learn through images, classical methods require numerical input. Features extracted through HOG tend to describe images through edge orientations in local regions of

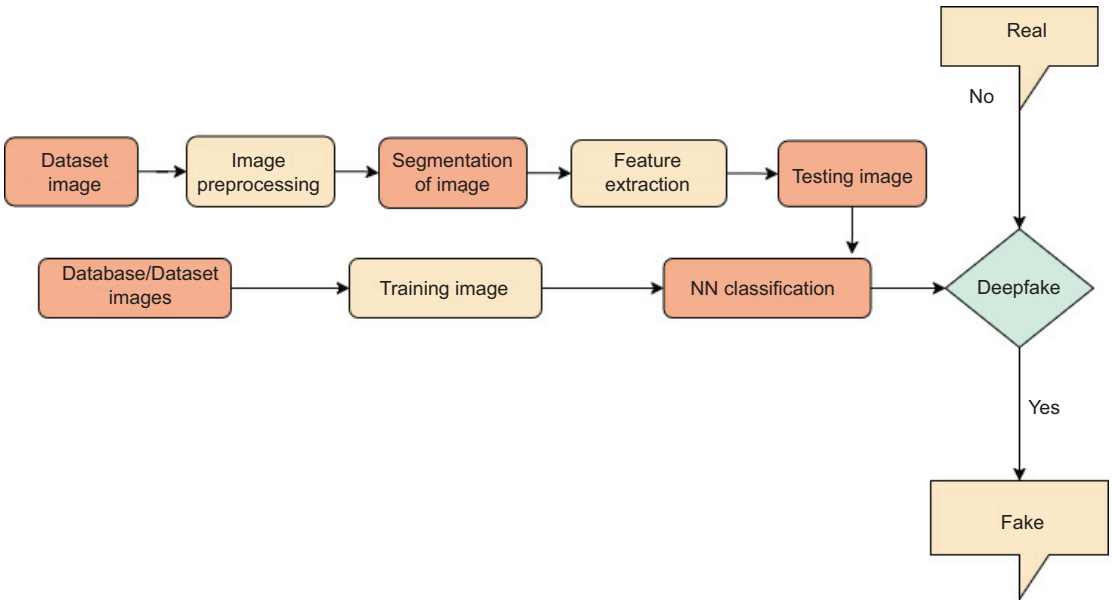

Figure 2. Work flow.

facial images, which tend to vary upon manipulation of facial images.

Additionally, the images were converted to greyscale and downsampled to give them consistent resolution before extracting features from them. Feature vectors from each image, based on the extracted HOG features, were used to enhance images and to normalise them using a standard scalar. These normalised vectors formed the input to the traditional classifiers used in this study.

Four ML models – SVM, Decision Tree, Random Forest, and gradient boosting – are trained using the extracted features by applying the HOG descriptor method. These ML models cover a range of ML paradigms to include margin methods, tree methods, etc. Their comparative evaluation resulted in gaining valuable insights into how different classical models make use of handcrafted features in terms of deepfake detection.

2. *Automated feature learning*: Apart from manual approaches to feature learning, DL approaches to automatic feature learning were used with CNNs. While other approaches rely on various feature representations, CNN systematically learns discriminative features directly from raw image pixels. It has to be mentioned here that CNN models were trained on data after preprocessing, wherein all images were resized into a certain spatial dimension, either 64 × 64 or 128 × 128. Moreover, pixel values were normalised to a specified range (0, 1). For these

models, multiple convolutional and pooling layers were put together with fully connected layers.

By means of the hierarchical learning approach, the proposed CNN was able to capture feature representations at various levels of abstractions, from edge information to higher- level texture and structure information pertaining to faces. Such feature learning capabilities and intra-manipulation recognition potential of the model explain its ability to recognise manipulation artifacts not easy to incorporate through feature engineering. The learning mechanism also presented a baseline model for the evaluation of performance variations by means of DL compared with ML.

5. Classical Machine Learning Models

Classical ML models were employed as baselines, which were originally based on classification of real and deepfake images, relying upon handcrafted HOG features. These classical ML models were chosen from a range of classification techniques, which vary in high interpretability, simplicity, and robustness. The studied ML model classification techniques include SVM, Decision Tree, Random Forest, and gradient boosting. Each model was trained on a common feature representation, thus ensuring a good comparative basis. Hyperparameters were adjusted with validation data to minimise overfitting. The results obtained from these models help in understanding how well are traditional learning models in performing deep fake detection under feature representation constraints.

5.1. Support Vector Machine

The SVM has been tested because it is very effective in binary classification and allows learning the optimal boundary in high-dimensional feature space. The HOG-based feature vectors learned a separating hyperplane for real versus manipulated images using SVM.

Both linear and non-linear kernels were considered, but the best performance was given by the RBF kernel. Since this is a non-linear kernel, there is an extra regularisation parameter C. This balanced margin maximisation with classification accuracy, and C was tuned. SVM has shown efficient training and strong baseline performance; however, due to its reliance on handcrafted features, it misses out on the chance of capturing higher-level facial representations, as opposed to DL methods.

5.2. Decision Tree

Choosing to evaluate the Decision Tree classifier is based on its ability to present decisions in an easily interpretable form. With the HOG feature set as input to the classifier, hierarchical decisions based on gradient/edge detection were learned by the classifier. Despite achieving high accuracy in training, the model showed signs of overfitting, especially after eliminating constraints involving tree growth. Constraints involving a maximum tree limit, a limit of samples in each node, were introduced. The Decision Tree model, even though unsuccessful in outperforming the ensemble-based model, helped in understanding more about DL in a deepfake image classification scenario.

5.3. Random Forest

Random Forest was used to overcome issues of instability and overfitting in a single decision tree. For example, predictions from a collection of decision trees are combined by Random Forest to improve robustness. The model was trained with varying number of trees and showed improvements over using a single Decision Tree algorithm. The feature importance test also showed what features of HOG-based features were most important in influencing decisions. The Random Forest algorithm also showed promising capabilities of classical models with regard to precision and reduction of variance.

5.4. Gradient Boosting

Gradient boosting was assessed as a sequential ensemble approach to enhance classification performance incrementally by concentrating on misclassified data from the previous step. Gradient boosting produced best performance on the same set of HOG feature representations as other classical ML models. Despite having more training time, compared to the model used in the Random Forest algorithm, precision and recall abilities were enhanced by the model's relentless efforts in refining decision boundaries for classification. This showcases the perfect applicability of boosting techniques utilised by the model in dealing with the slight facial changes introduced by deepfake techniques.

6. Deep Learning Models

Deep learning techniques were used to mitigate the drawbacks of handcrafted feature extraction by utilising the potential of machines to learn directly from image data. CNNs were chosen, as they are successful in image-based pattern recognition.

A major feature of CNNs is that they learn hierarchical feature representation from pixel values. This feature eliminates any need to add artificial features. Therefore, it creates a capacity in the model to identify complex spatial relationships and artifacts, including those created in deepfake images of faces. The CNN structure developed has representational and computation efficiency. As a result, it presents a lightweight model for deepfake detection.

6.1. Convolutional Neural Networks

A CNN was applied to carry out the task of binary classification on real and deepfake images. It comprised various layers where features are extracted through convolution and max pooling operations, followed by classification through fully connected layers. The input is resized to various sizes and passed through CNN to obtain corresponding features and output classes. This is followed by normalising of features, and then CNN was designed based on three dimensions: 64 × 64. Finally, convolutional layers learned the spatial features of increasing complexity layer by layer. Accordingly, pooled layers simplified the output of convolutional layers. To decrease the model's tendency to overfit the dataset provided for training – due to its relative scarcity – a dropout layer was also incorporated in the network's architecture. On the other hand, the output layer of the network used the sigmoid activation function. This CNN architecture allowed for the automated extraction of discriminative features from facial areas and showed improved results over traditional model approaches, which depended upon handcrafted features, with low computational complexity. State-of-the-art architectures, such as the Xception network, the EfficientNets family, along with ResNets, are omitted in the present work because the objective of this study was not to outperform existing state-of-the-art benchmarks, but to compare classical machine learning approaches with a lightweight CNN under controlled experimental settings. Instead, it is to determine the performance gap between traditional ML approaches and a bare-bones deep learning configuration.

Table 3 shows the optimized hyperparameter settings used for the CNN architecture.

7. Evaluation Metrics

Following the training and validation process, we compared our models to a number of recognised measures, including F1-score, accuracy, precision, recall, and area under the curve–receiver operating characteristic (AUC-ROC). We evaluated the

Table 3. Optimized CNN hyperparameter settings.

Hyperparameter	Value	Description
conv_blocks	2	Number of Conv+pooling blocks
filters_0	64	Filters in first conv layer
kernel_size_0	3	Kernel size for convolution
dense_units	192	Units in fully connected layer
dropout	0.4	Dropout rate for regularization
learning_rate	0.00064094	Learning rate for optimizer

models' overall ability to generate forecasts using more intricate ways by evaluating them in opposition to a variety of measures. This gives us the basis for deeper assessments of the systems' practical performance.

7.1. Accuracy

Accuracy is the proportion of correctly classified observations to total observations. This is a crude indicator of the performance of our models. Although accuracy is a rapid overview of the performance of a model, it can be misleading with imbalanced datasets. A model that consistently predicts 'true' attains an accuracy rate of 90% in predicting correct observations, but it yields no useful information if, for example, 90% of the photos are true and 10% are fake.

7.2. Precision

Knowing how many of the identified images based on detection are actual deepfakes is important for precision. Precision is determined by dividing true positives by all predicted positives. If precision is high, the false positive rate is low. This is especially important for delicate applications, because labelling a genuine image in the media as a fake might have practical repercussions.

7.3. Recall

In contrast, recall measures how many of the actual deepfake images are true positives identified by the model. Recall is the fraction of true positives to the total actual positives. A model with high recall is one that has identified many deepfakes, but must also have a high false positive rate. We must strike the right mix between recall and precision for our purpose.

7.4. F1-score

F1-score is the harmonic mean of precision and recall, giving us one number that fitted these two metrics together to report in place of precision and recall. This is beneficial when there is class imbalance in the dataset to ensure we are not missing valuation of one of those metrics or the other when evaluating the model. In a deepfake detection problem, the cost of accidentally missing a fake and misclassifying a real image has essentially the same consequences and importance to a practitioner and thus provide F1-score as a balanced perspective.

8. Classical Machine Learning model

Support vector machine, Decision Tree, Random Forest, and gradient boosting are four different predictive models that went through training to catch deepfakes. Each model was fine-tuned and later evaluated based on related metrics.

8.1. Support vector machine

The SVM classifier is designed to make a hyperplane that separates real and fake images in the feature space. SVM demonstrated its ability to define a choice border in our research by achieving an appropriate precision and recall result. However, due to its sensitivity to noisy feature, its performance dropped on some image classes, which is why we had a somewhat moderate F1-score.

8.2. Decision Tree

The output of the Decision Tree classifier has good prediction classification accuracy, although remaining interpretable. With some of our initial evaluations, we noted evidence of overfitting to the training data, although they did yield training accuracy results that were high with accuracy and F1-scores slightly lower in the testing phase. Testing round models may also indicate confusion, with borderline cases contributing to form noise in AUC measure of visible performance assessment.

8.3. Random Forest

Random Forest made major advancements towards solving Decision Tree destabilisation through its collective approach. Both precision and recall of the model were higher, showing some level of resistance to overfitting. The AUC-ROC measure for Random Forest was also among the best of all the classical models

evaluated, suggesting the model's considerable effectiveness for distinguishing between deepfake and real images across decision thresholds constrained by the analysis.

8.4. Gradient boosting

The gradient boosting enhanced effectiveness by successfully reducing mistakes in classification. This model produced the highest F1-score among purely classical models based on the highest precision and recall. This improved performance with high AUC-ROC value confirmation, indicating a better-performing approach, especially when it came to identifying smaller size deviations in affected photos. Figure 3 compares the performance of the evaluated classical machine learning models.

9. Results and Discussion

The accuracy of 93% proves relevant only in a specific context of experimental scope. Various aspects contribute towards this performance metric, even in a relatively small dataset. First, the dataset is relatively balanced in terms of real and artificial samples, thus eliminating classification bias. Second, as a preprocessed dataset, face-based processing trains a model that ignores irrelevant background noise while focusing only on artificial facial regions. Lastly, testing a model in a dataset from a similar population as that of training results in better performance metrics in this experiment

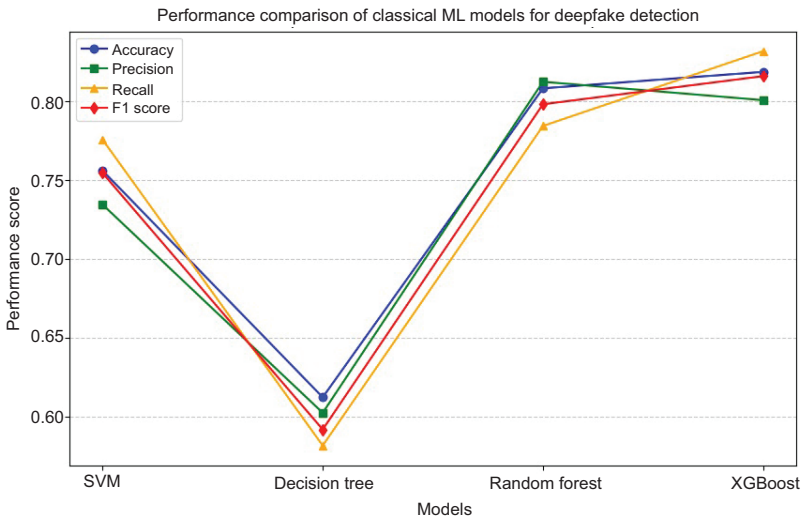

Figure 3. Performance comparison of classical ML models.

compared to cross-dataset testing. Interestingly, it should be noted that this achievement does not necessarily mean that the suggested models generalise well to other deepfake generative methods and realistic circumstances. Rather, it signifies that an even lightweight version of CNNs might well identify discriminative facial features through restricted and well-controllable circumstances.

Figure 4 shows the accuracy comparison among the evaluated models. Figure 5 presents the precision comparison among different models. Figure 6 illustrates the recall comparison of the evaluated models. Figure 7 presents the F1-score comparison among the evaluated models.

Table 4 presents the comparative performance of all evaluated models using accuracy, precision, recall, and F1-score metrics.

As shown in Table 4, CNNs achieved the highest performance across all evaluation metrics. To further examine the classification effectiveness of the proposed model, additional evaluation techniques, including a confusion matrix and ROC curve, were utilised.

Figure 8 shows evaluation of a CNN model on the tested dataset. The confusion matrix is a detailed representation of the classification accuracy of the model on real and deepfake images. A total of 1682 real images were classified correctly as real, and 1214

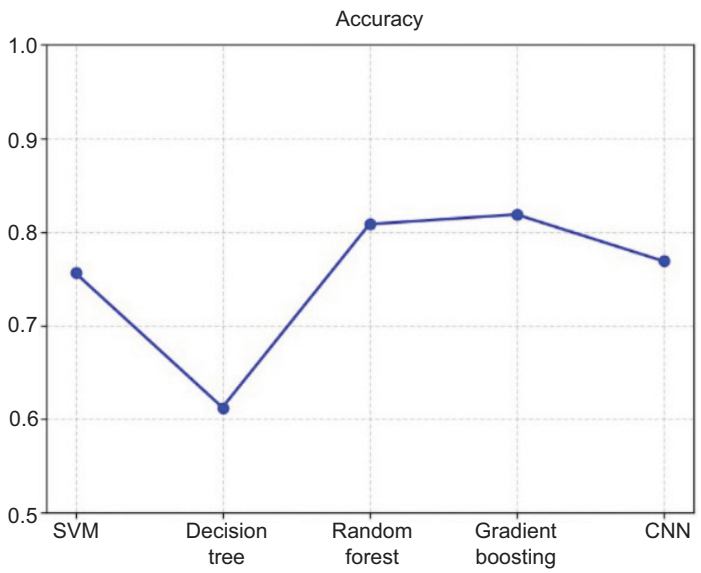

Figure 4. Model accuracy comparison.

Table 4. Model performance comparison.

Model	Accuracy	Precision	Recall	F1-score
SVM	0.84±0.02	0.82±0.02	0.80±0.02	0.81±0.02
Decision tree	0.81±0.03	0.78±0.02	0.77±0.03	0.77±0.01
Random forest	0.89±0.01	0.88±0.01	0.87±0.01	0.87±0.02
Gradient boosting	0.91±0.01	0.90±0.01	0.89±0.01	0.89±0.01
CNN	0.93±0.01	0.94±0.01	0.92±0.01	0.93±0.01

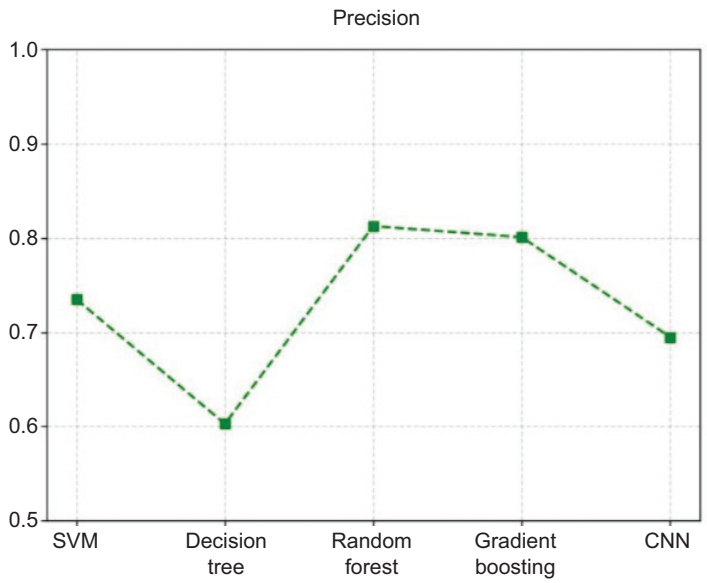

Figure 5. Model precision comparison.

deepfake images were classified correctly as deepfake. This shows that the proposed model has a strong ability to detect real facial content as well as synthetic modifications. However, the confusion matrix also shows that 278 real images were classified incorrectly as deepfake images, and 756 deepfake images were classified incorrectly as real images. This shows that although the proposed model has a high classification accuracy, some deepfake images have characteristics that are very similar to real images, making them more difficult to detect. The dominance of diagonal elements of the confusion matrix shows that the proposed model has been able to learn effective feature representations, thereby confirming that the CNN model has been able to learn discriminative features for deepfake image detection. The confusion matrix is an indication

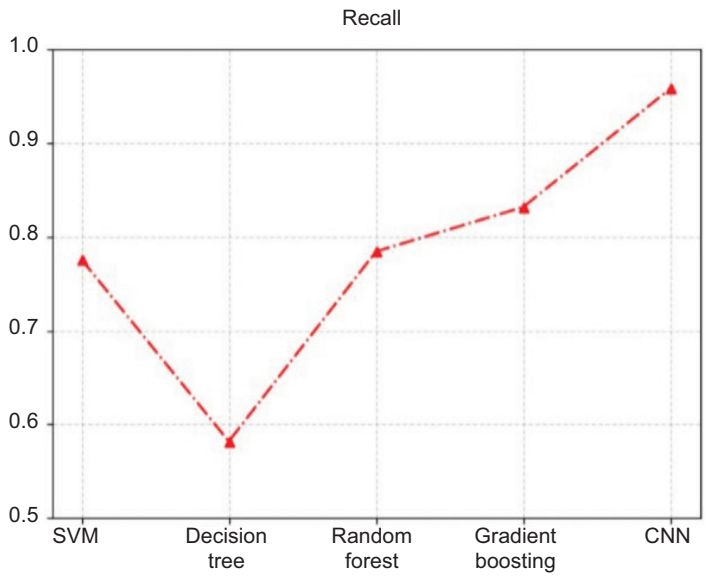

Figure 6. Model recall comparison.

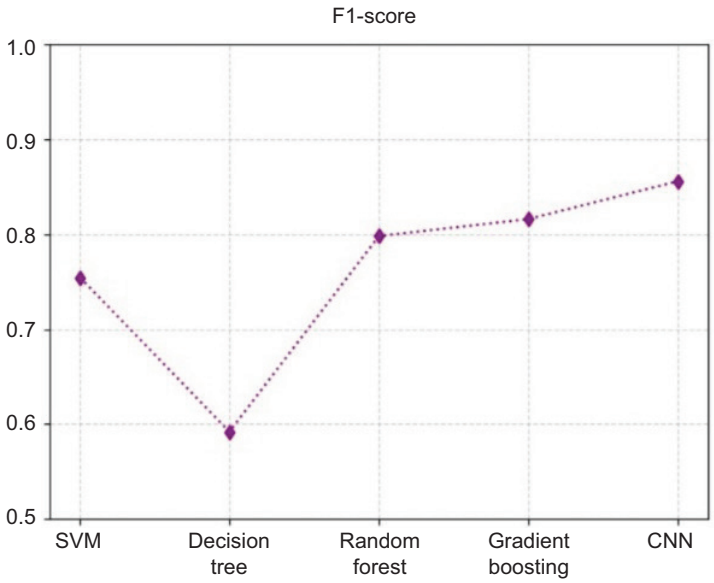

Figure 7. F1-score comparison among the evaluated models.

that the proposed model is robust but needs improvement in the reduction of false classifications.

Figure 9 is the ROC curve of CNN model of test data. The ROC curve is a plot of true positive rate against false positive rate for different

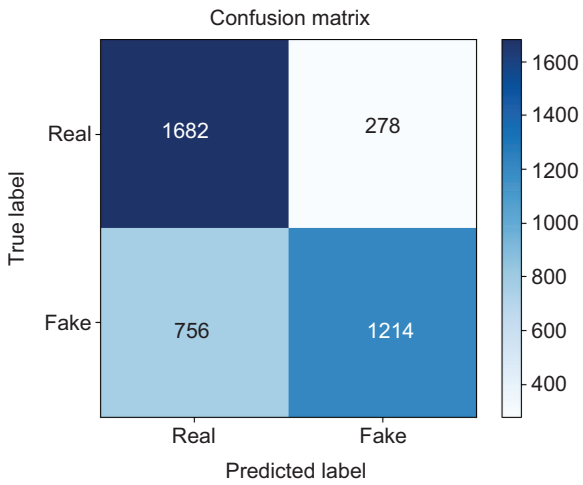

Figure 8. Confusion matrix showing the classification results of CNN model on test dataset.

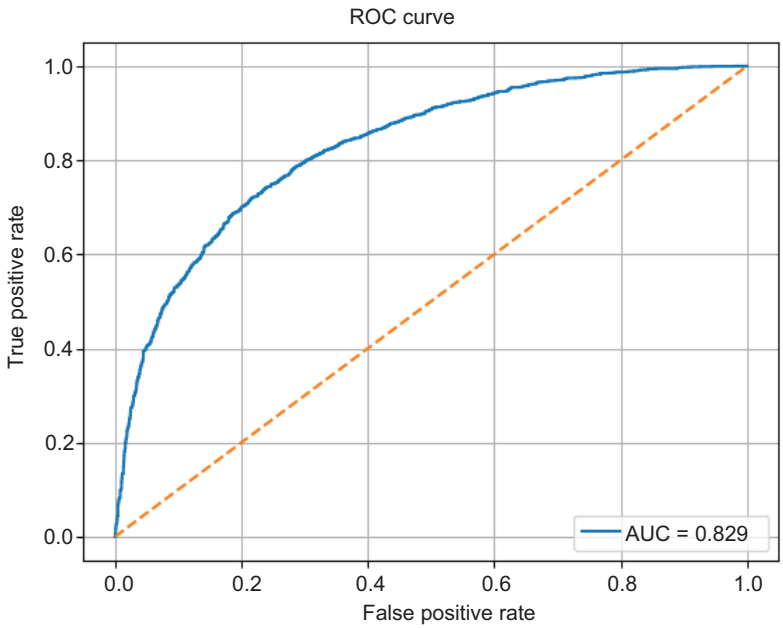

Figure 9. Receiver operating characteristic curve illustrating the discriminative capability of a CNN model.

threshold settings. The AUC score of the model is 0.829, which has a good discriminative ability to distinguish between real and deepfake images. An AUC value far greater than 0.5 ensures that the model performs significantly better than a random classifier. Smooth curve with a positive slope indicates that the model has a

good balance between sensitivity and specificity, which is capable of detecting manipulated images as well as suppressing false alarms. These results confirm the efficacy of the proposed DL approach for deepfake image detection in a controlled experimental setup.

Further improvements in feature extraction and model optimisation could enhance the detection capability of deepfake samples, as indicated by the AUC value. Results are reported as mean ± standard deviation across multiple experimental runs to ensure robustness and reproducibility.

10. Limitations and Future Work

This study has the following limitations since the study used fewer image-based models that were constructed on video datasets, thereby limiting extension to real-world applications. Cross-dataset testing and adversarial testing for robustness was not conducted and need to be further explored in future. Limited model architectures were used for testing in the study; thus, future studies need to include more model architectures, since the study used fewer models for testing, while the rest of the models need to be explored for their applicability to the image classification task and more than just the single modal approach to the task. The study is conducted on a deliberately reduced dataset (3930 images), and therefore findings are more applicable to controlled small-data scenarios rather than unconstrained real-world settings.

11. Conclusions

In order to detect deepfake films, we compared DL and classical ML techniques in this research. We specifically looked at how well each method performed using the FaceForensics++ dataset. The prevalence of deepfake technology has led to a considerable increase in worries over the validity of digital information. Our study's main objective was to compare the accuracy and generalisability of DL techniques, like CNNs, and classical ML techniques, like SVM, Decision Trees, Random Forests, and gradient boosting.

Our research showed that although traditional ML models are easier to understand and use less computing power, they are not as accurate as DL-based methods for detecting deepfakes. CNNs consistently beat models like SVM and Decision Trees across all assessment measures, even if they performed quite well after being trained on features like HOG.

Our study's most intriguing finding was how much the feature extraction technique selection affects the performance of traditional ML models. HOG features performed well in collecting important face aspects in our tests, but they were unable to handle intricate, high-level patterns that DL models are capable of acquiring automatically. This was particularly true for CNNs, which can automatically extract complex characteristics from raw video frames because of their layered design, providing a considerably better comprehension of the input material.

Additionally, we experimented with a hybrid strategy that blended CNNs with conventional ML models, such as Random Forest and XGBoost. This strategy showed promise as it allowed us to take advantage of CNNs' powerful feature extraction capabilities while maintaining the classification strength of traditional ML models. DL-only models, however, continued to beat hybrid strategy, highlighting the superiority of DL approaches for intricate tasks, such as identifying deepfakes.

The amount of training data was another important consideration in our research. Our findings demonstrated that more data improves the performance of both DL and traditional models. The accuracy improved as the training set size grew, which is in line with the results of other studies. This implies that deepfake detection methods, particularly DL-based models, need a lot of labelled data to function well.

Although DL models are obviously superior, there are still some evident benefits of using traditional ML techniques. They are perfect for settings with constrained computer capacity since they are more resource-efficient and need less computational resources to train. Better interpretability is another benefit they offer, which is crucial for situations where it's essential to comprehend how a model arrives at its conclusions. Classical models can be useful for rapid deployment in resource-constrained environments if they are trained on well-designed feature sets.

There are a number of intriguing avenues for advancing deepfake detection systems in future. To strengthen detection algorithms, integrating multimodal data – such as video frames with audio or metadata – is one possible strategy. The quantity of training data needed for DL models may be decreased with the aid of research into transfer learning techniques, increasing their usability and effectiveness. The development of real-time detection algorithms that recognise even the most complex deepfakes is crucial as deepfake technology advances.

To sum up, our study shows the advantages and disadvantages of DL and traditional ML methods for identifying deepfakes. DL models – particularly CNNs – consistently outperform traditional ML models in terms of accuracy and resilience, despite the latter's simplicity and efficiency. DL and other sophisticated approaches are becoming increasingly important as deepfake technology advances. Nevertheless, traditional models still have a role, especially where computing resources are scarce or model interpretability is crucial. The results of this study aid in the continuous creation of deepfake detection systems that are more effective and efficient, and are crucial for maintaining the integrity of digital information in today's quickly changing technological environment.

Conflict of Interest

The authors declared that they had no conflict of interest related to the publication of this manuscript.

Data Availability Statement

The datasets used and analysed in the current study are available publicly. Specifically, the FaceForensics++ and Celeb-DF datasets utilised in this research can be accessed, respectively, at https://github.com/ondyari/FaceForensics; https://github.com/yuezunli/celebdeepfakeforensics.

References

[1] N. Sandotra, B. Arora, "A comprehensive evaluation of feature-based AI techniques for deepfake detection," *Neural Computing and Applications*, vol. 36, no. 8, pp. 3859–3887, 2023, doi: 10.1007/s00521-023-09288-0.

[2] O.A. Shaaban, R. Yildirim, A.A. Alguttar, "Audio deepfake approaches," *IEEE Access*, vol. 11, pp. 133000–133020, 2023, doi: 10.1109/ACCESS.2023.3333866.

[3] L. Stroebel, M. Llewellyn, T. Hartley, T. Shan Ip, M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *Journal of Information Technology Case and Application Research*, vol. 25, no. 2, pp. 95–118, 2023, doi: 10.1080/23742917.2023.2192888.

[4] Md. S. Rana, A.H. Sung, "Advanced deepfake detection using machine learning algorithms: A statistical analysis and performance comparison," in *Proceedings of the International Conference on Intelligent Computing and Information Technology (ICICT),* pp. 1–6, 2024, doi: 10.1109/ICICT62343.2024.00019.

[5] M.U.T. Gujjar, K. Munir, M. Amjad, A. Ur Rehman, A. Bermak, "Unmasking the fake: Machine learning approach for deepfake voice detection," *IEEE Access*, vol. 12, pp. 14522–14540, 2024, doi: 10.1109/ACCESS.2024.3521026.

[6] M. Taeb, H. Chi, "Comparison of deepfake detection techniques through deep learning," *Journal of Cybersecurity and Privacy*, vol. 2, no. 1, pp. 7–19, 2022, doi: 10.3390/jcp2010007.

[7] A. Heidari, N.J. Navimipour, H. Dag, S. Talebi, M. Unal, "A novel blockchain-based deepfake detection method using federated and deep learning models," *Cognitive Computation*, vol. 16, pp. 1150–1167, 2024, doi: 10.1007/s12559-024-10255-7.

[8] M. Dang, T.N. Nguyen, "Digital face manipulation creation and detection: A systematic review," *Electronics*, vol. 12, no. 16, Art. no. 3407, 2023, doi: 10.3390/electronics12163407.

[9] K.A.P. da Costa, D. Jodas, L.A. Souza Jr., "A review of deep learning-based approaches for deepfake content detection," *AI*, vol. 4, no. 3, pp. 410–430, 2023, doi: 10.3390/ai4030021.

[10] A. Tiwari, R. Dave, and M. Vanamala, "Leveraging Deep Learning Approaches for Deepfake Detection: A Review," in Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI '23), pp. 12–19, 2023, doi: 10.1145/3596947.3596959.

[11] N.Y. Kapadnis, A. Rajesh, "Deep fake detection using CNN," in *Proceedings of the International Conference on Computing Systems and Applications*, pp. 112–118, 2024, doi: 10.1109/ICCSA60280.2024.10499999.

[12] N. Sandotra, B. Arora, "Deep learning-based model for deepfake image detection: An analytical approach," *Multimedia Tools and Applications*, vol. 83, pp. 22451–22470, 2024, doi: 10.1007/s11042-024-18521-0.

[13] L. Wang, D. Li, X. Meng, "Deepfaker: A unified evaluation platform for facial deepfake and detection models," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4550–4565, 2024, doi: 10.1109/TIFS.2024.3376543.

[14] Y. Patel, S. Tanwar, R. Gupta, "Deepfake generation and detection: Case study and challenges," *Multimedia Tools and Applications*, vol. 82, pp. 31211–31239, 2023, doi: 10.1007/s11042-023-15012-5.

[15] P. Edwards, J.-C. Nebel, D. Greenhill, X. Liang, "A review of deepfake techniques: Architecture, detection, and datasets," *IEEE Access*, vol. 12, pp. 22010–22045, 2024, doi: 10.1109/ACCESS.2024.3367812.

[16] A. Heidari, N.J. Navimipour, H. Dag, M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Artificial Intelligence Review*, vol. 56, pp. 10567–10612, 2023, doi: 10.1007/s10462-023-10475-6.

[17] M.U.T. Gujjar, K. Munir, M. Amjad, A. Ur Rehman, A. Bermak, "An investigation into the utilisation of CNN with LSTM for video deepfake detection," *IEEE Access*, vol. 12, pp. 90211–90229, 2024, doi: 10.1109/ACCESS.2024.3412345.

[18] L. Whittaker, R. Mulcahy, K. Letheren, J. Kietzmann, R. Russell-Bennett, "Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda," *Technological Forecasting and Social Change*, vol. 197, Art. no. 122901, 2023, doi: 10.1016/j.techfore.2023.122901.

[19] S.H. Al-Khazraji, H.H. Saleh, "Impact of deepfake technology on social media: Detection, misinformation, and societal implications," *Journal of Information Security and Applications*, vol. 75, Art. no. 103485, 2023, doi: 10.1016/j.jisa.2023.103485.

[20] S. Sadiq, T. Aljrees, S. Ullah, "Deepfake detection on social media: Leveraging deep learning and fasttext embeddings for identifying machine-generated tweets," *Expert Systems with Applications*, vol. 232, Art. no. 120812, 2023, doi: 10.1016/j.eswa.2023.120812.