A Hybrid CNN-Ensemble Framework for Robust DeepFake Image Detection

Mohammad Alsulami
College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia

***Abstract*—The fast development of deepfake technologies has caused growing concerns related to the authentication of digital media, the integrity of personal identification, and the spreading of disinformation. Therefore, there is a growing need for effective automatic detectors of deepfakes. Meanwhile, existing techniques of deepfake detection encounter a large number of difficulties. First, it is difficult to distinguish the subtle manipulation details of faces. Another problem is poor generalization when applying models to novel datasets or highly realistic, generated synthetic faces. Another issue is low accuracy in the case of imbalanced data, where samples of one class dominate others. Thus, to address those problems, this study proposes a novel hybrid approach based on a combination of deep learning (DL) and conventional machine learning (ML) for detecting deepfake images. More precisely, two pre-trained CNNs (MobileNetV2 and ResNet50) were applied to generate features and classify them via the Random Forest (RF) algorithm. The experiments have been conducted on a benchmark of 6,557 facial images marked as either real or fake. The findings reveal that the MobileNetV2+RF achieved the highest accuracy 99%, followed by MobileNetV2 with 98% and ResNet50 with 97% accuracy. This suggests that the hybrid architecture helps to increase the effectiveness of the solution. A statistical significance test reveals that none of the models' performances differ significantly from each other (p > 0.05). Overall, the proposed system demonstrates excellent metrics concerning accuracy and precision.**

Keywords*—*Deep learning models; machine learning classifiers; image processing; data augmentation

I. Introduction

The rapid evolution in the fields of artificial intelligence (AI), deep learning, and multimedia technologies has transformed the way people interact and engage in creating digital information [1]. Currently, smartphones, laptop computers, and sophisticated computer systems provide individuals with the capacity to develop, disseminate, and access digital multimedia in a faster way by using online platforms. This technological development has enabled many benefits in fields such as entertainment, the educational sector, health care services, and even social media communications. However, they have posed many problems associated with digital security and integrity [2]. Deepfake technology is one of those serious problems.

Deepfakes refer to media that have been modified using digitally manipulated images, videos, or audio that were produced through artificial intelligence techniques [3]. The artificial intelligence used includes generative adversarial networks and autoencoders. Through this technology, highly accurate fake faces and voices have been produced since the technology is capable of mimicking human facial expressions and their voices. Deepfake technology has experienced great progress ever since its invention in 2017 due to the emergence of free and available artificial intelligence technology, the availability of big data online, and the availability of computational power [4].

Misuse of deepfake technology has caused several significant threats to individuals, organizations, and even society as a whole. The uses of deepfakes have been identified to mislead people, shape their opinions politically, engage in identity fraud, conduct financial crimes, and create inappropriate content. Social media platforms like Facebook, Instagram, Youtube have made deepfake videos and other forms of content go viral on the internet in an instant due to the speed at which information can travel over them. On the other hand, deepfake technology is a concern in many fields, including forensics, surveillance, and biometrics [5], [6], [7].

Facial recognition and identification systems based on conventional methods depend heavily on using biometric data associated with the face. In most cases, such algorithms do not cope with the identification of manipulated images of people's faces. They are less efficient with images that have undergone various types of manipulations, such as blur processing, compression, sizing, filtering, and enhancement. Meanwhile, development in the field of image manipulation and AI-based face generation has increased the realism of synthetic faces [8], [9].

To address these issues, recent works in this field have been concentrating on designing techniques for automatically detecting deepfakes through deep learning methods. Convolutional Neural Networks (CNNs) have been mostly applied since they are able to efficiently extract visual features from these manipulated images. Models like MobileNetV2 and ResNet50 are usually used because of their excellent ability to do so, along with being computationally efficient. Nevertheless, it might not be effective to solely rely on deep learning-based techniques for obtaining the optimal results [10] [11].

Recent studies reveal that a combination of deep learning methods with classifiers of machine learning is useful for enhancing detection precision and reliability of the overall framework. Methods of machine learning, such as Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), can be used to make decisions using features obtained from CNN algorithms. Hybrid models tend to increase efficiency, feature exploitation, and generalization power over pure deep learning-based frameworks [12].

Even though advancements have been made to improve deepfake generation and detection systems recently, current facial forgery detectors still experience certain limitations while working with sophisticated manipulations within realistic images. In particular, traditional approaches demonstrate decreased effectiveness in the presence of manipulations performed by means of compression, resizing, filtration, and blurring. Moreover, several current approaches require the use of a deep learning classifier only, meaning the high level of computation costs and inability to work efficiently with diverse manipulated images. Conversely, the fast development of AI-based facial image manipulation has led to the necessity of creating new detection systems, allowing for achieving both effective feature extraction and efficient classification in order to ensure higher levels of performance and detection effectiveness. Thus, there is a significant need for developing an approach for deepfake detection that would involve both deep learning and machine learning algorithms [13], [14].

The reason behind developing a dependable, effective, and generalized model for detecting deepfakes became one of the key areas of study in digital security and forensics. Reliable frameworks are needed to maintain confidence in digital media, develop biometric recognition systems, decrease the dissemination of misinformation, and verify the authenticity of digital data.

The primary goal of this research work is to design a detection model utilizing deep learning and machine learning techniques. In the proposed approach, pre-trained CNNs would be used to extract the features from the images. Machine learning classifiers would be utilized to increase the reliability and performance of the detection algorithm. To achieve our objective, the proposed study was designed following the research questions (RQ):

RQ1: How effectively do transfer learning-based CNN models like MobileNetV2 and ResNet50 help in detecting important facial characteristics to detect deepfake images?

RQ2: Is there any improvement in the detection of deepfake images if we combine ML classifiers such as Random Forest (RF) with deep learning-based facial feature extraction methods?

RQ3: Which is the most suitable and accurate combination of deep learning models and ML classifiers for detecting deepfake facial images?

Some of the key contributions of this research work are as follows:

- A novel framework for detecting deepfakes by implementing machine learning and deep learning approaches has been introduced.
- Deepfakes have been detected by using transfer learning-based CNN architectures such as MobileNetV2 and ResNet50.
- Different machine learning classifiers like RF, SVM, and logistic regression have been utilized to increase classification performance.
- Data augmentation, class imbalance problem, early stopping, and learning rate scheduling have been utilized to enhance generalization performance.
- Experiments were performed based on the ROC curve, confusion matrix, and comparative analysis of accuracy.

This study is significant since it helps in solving one of the biggest issues currently facing digital security: identifying manipulated facial content using Artificial Intelligence. This study presents some contributions to cybersecurity, biometrics, digital forensics, and media validation through the improvement of AI deepfake detection technology. This study is practically applicable to social media monitoring, online identity verification systems, surveillance systems, and misinformation detection, among others. Additionally, the use of deep learning and machine learning in this process gives a precise and effective computational method to help design intelligent security systems in the future.

The rest of the study is structured as follows: Section II identifies the key research gap. In Section III, detailed steps for conducting the proposed study are discussed. Key findings, interpretation, and discussion are explained in Section IV. Section V outcome / future directions are presented.

II. Literature Review

Deepfakes are facial images that are generated using AI. It is a major problem for the security of the web, privacy online, and the authenticity of data present online. Several deep learning architectures can be used to detect DeepFakes. There are several pretrained CNN architectures, such as Inception netv3, Resnet 50, Resnet 101, and DenseNet121, trained using available benchmark datasets such as real and fake face detection using 2k samples, real versus fake face detection using 140k samples, and deepfake and real images using 190k samples. The efficiency of these pretrained CNN models depends on the availability of the number of samples and their resolution. The current pretrained models do not perform well in terms of efficiency; for example, a pretrained model performs with an accuracy of 55% in an imbalanced and limited data set, such as the one with a real and fake data set of 2k samples. In this study, a hybrid DL model combining both the capability of a pretrained model and a GAN with an SE attention mechanism achieves an accuracy of 89% [15]. With the rising cases of counterfeit digital images and increasing adoption of the deep fake technology, there is a need for developing efficient algorithms that can be able to detect counterfeit pictures. This study [16] provides a deep neural network algorithm known as DeepFaceGuard with an objective of detecting counterfeit pictures using a multi-layer architecture in order to ensure low detection time without compromising on accuracy. A pre-trained model of VGG-16 was used to increase visualization, while the principal component analysis technique was used to reduce dimensions. Accuracy and F1-score of 0.83 were some of the measures used to evaluate the proposed model by use of 140,000 image dataset. It was found that DeepFaceGuard had an accuracy of more than 83%. This shows its efficiency in detecting counterfeit images. DeepFaceGuard attained a verification accuracy of 84% in only 184 seconds (8 seconds × 23 epochs). Prior studies involved residual convolutional blocks or the use of an attention mechanism for DF detection; unfortunately, most researchers consider each of these separately without any emphasis on fairness, out-of-distribution generalization, statistical tests, or artifact-based detection. This study [17] proposes a method named ARC-Net that employs both attention and residual convolutional blocks, along with a popular base, EfficientNet B0, and an attention mechanism. For testing ARC-Net alongside the rest of the four pre-trained DL models, a real image dataset of 500 images from Bangladesh, in addition to a 140k mixed real and fake faces data set, was used. ARC-Net outperformed the traditional and state-of-the-art methods significantly with an accuracy of 99%, a precision of 1.0, a recall of 0.97, and an F1 score of 0.98, obtaining the best possible reliability for DF detection. For evaluating the external validity and generalizability of ARC-Net, two new data sets, the Deepfake dataset and the Deepfake database, have been tested for the model, and it showed consistent and highly reliable results in all experiments conducted, which also include varying scales. Three OOD tests were performed; the first test examined the South Asian dataset, and we observed that including less than one percent of the actual Bangladeshi image data lowered the FPR and enhanced the probability calibration significantly, whereas the other two cross-dataset tests exhibited high transferability. We have also done an ablation test to observe the effect of various components in a model by deleting and changing them, and we have used McNemar's Statistical Test to determine statistical significance between rival classifiers. Furthermore, we have utilized Explainable AI (XAI) approaches such as Grad-CAM and LIME to provide transparency regarding the outcomes, considering that the facial region needs to be prioritized for detection. The fast-paced development of Deep Learning (DL) algorithms has caused the extensive emergence of DeepFake (DF) images in real life, which raises real-life challenges. In the present work, DeepDect is proposed and evaluated within a real-life context. The platform has been developed through a human-centered approach, considering insights (requirements) from both regular users and DF experts. The most effective DL models (ResNet-50 and Random Forest for face-swapping detection, Capsule Forensics v2, and CNN for AI-generated images) have been incorporated into the platform as a detection engine. An Explainable AI (XAI) module has been designed and incorporated into DeepDect for visual (Grad-CAM heatmaps) and textual explanations. Real-life evaluation has been conducted with the participation of 108 users to measure the performance of DeepDect against humans. DeepDect has obtained 81% detection accuracy and outperformed human users in terms of DF detection rate, emphasizing the necessity for such a solution in real-life contexts [18]. With the fast development of face manipulation techniques through Generative Adversarial Networks (GANs) and Stable Diffusion-based approaches, this study [19] presents an investigation into the topic of deepfake content generation. This study examines the possibility of utilizing Vision Transformers (ViTs) for deepfake image detection based on the ability of such networks to extract global features. To determine the efficiency of ViTs for multiclass deepfake image detection was used. The proposed model shows high levels of accuracy, precision, and recall, and the F1 score of 99.90% on a multiclass-prepared dataset. The presented findings show that there is a great potential in ViTs to help make the internet safer from deepfake threats, especially caused by Stable Diffusion and StyleGAN2. The developed model demonstrated better performance than state-of-the-art CNN-based models, i.e., ResNet-50 and VGG-16. Currently, deepfakes have been considered a popular research area because of the considerable improvements made in the field of artificial intelligence (AI). Images produced by the generative AI system are realistic and entirely new. Hence, detecting the fake material using AI is very important. In this study[20], transfer learning was implemented on four models, including VGG16, VGG19, MobileNetV2, and ResNet50V2. The proposed model in this study is the one that is developed based on the pre-trained VGG16 model. In this method, we used the input of the VGG16 model along with the output of two more convolutional and MaxPooling layers for further feature extraction after concatenation, followed by three dense layers. According to the experimental values provided in this study, the proposed model achieves better performance compared to other approaches with 98.4% accuracy rate. It can be concluded that the detection of deepfakes will be feasible in the future through machine learning techniques. Deepfake technology development has outpaced that of video manipulation detection methods, posing some issues in terms of media security and forensic digital analysis. In this case study [21], we implemented a multimodel machine learning approach by implementing traditional as well as modern deep learning classifier models for deepfake video detection. For our case, the experiment made use of a subset of the dataset, which included 401 samples of training and 400 samples of testing data, but with an imbalance in classes in favor of FAKE videos. Our methods involved data preprocessing, EDA, face detection using Haar Cascade Classifiers in OpenCV, and data feature extraction. Models, including Random Forest, Support Vector Machines, and Convolutional Neural Network (CNN), among other models, were trained and tested using the model accuracy, precision, recall, and F1 score as metrics. Our deep learning model performed better than any other model in all metrics, indicating its capability in deepfake video detection. The study above shows how deepfake video detection requires a combination of classical and modern machine learning techniques. With the rise of hyper-realistic fake media generation techniques through deep learning applications, there is a need for detecting the authenticity of images online. In this study [22], an ensemble learning-based technique integrating InceptionV3, VGG16, and Xception architectures was applied to develop a deepfake detection system. The method yields an accuracy of 97.40% with 97.00% sensitivity and 98.00% specificity on a balanced dataset containing 70,000 real and 70,000 fake images. A comparison with ResNet50 and EfficientNet baseline models shows that our model is 5.6% more accurate than ResNet50 and 3.0% more accurate than EfficientNet in terms of AUC. Our findings indicate that the proposed method is promising in addressing the problem of detecting deepfake images. To assess the reliability of the developed model, extensive experiments were performed using a large number of real and synthetic images. In addition, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to interpret the decision of the model. The deepfake videos have been distributed in recent years using several deepfake techniques (i.e., faceswap, face2face, etc.). The aim of this study [23] was to design a model that helps detect deepfake videos. Two hybrid models, consisting of a combination of machine learning and deep learning, were proposed. The first hybrid model used CNN (convolutional neural network) for feature extraction along with different classifiers in machine learning, namely, SVM (Support Vector Machine), KNN (K-nearest Neighbor), DT (Decision Tree), RF (Random Forest), Logistic Regression, and Naive Bayes. In comparison, the second hybrid model utilized the concept of transfer learning using the VGG16 (Visual Geometry Group) pre-trained model and the same classifiers as the first model. Evaluation of both the hybrid models was done using the video dataset, FaceForensics++, containing four types of deepfake techniques: Deepfake, Faceswap, Face2Face, and Neural texture. Results obtained for both models were found to be promising due to high accuracy scores, proving the effectiveness of the proposed models that can be employed in the construction of a deepfake detector application. The first hybrid model gave maximum accuracy with the SVM classifier for the four deepfake techniques: 0.96, 0.87, 0.90, and 0.64, respectively. The impact of deepfake videos is immense and can range from identity theft to spreading misinformation to national security threats. Hence, there is a dire need to improve the performance of the algorithm used in deepfake detection. In this research [24], various feature extraction methods have been used to leverage the use of deep learning algorithms like Xception and ResNet50 in detecting deepfakes in a video using the DFDC data set. Further, several hybrid models comprising eight were created through the use of classification algorithms like SVM, KNN, MLP, and RF. ResNet50 and RF hybrid model showed the most promising results in terms of the accuracy rate, which stood at 98% and an area under the curve (AUC) of 99.65%. The emergence of manipulation on digital media poses a serious threat in today's society, hence the need to have accurate and reliable means of detection of these fakes. To provide an appropriate response to the problem, the study [25] proposes a novel hybrid structure known as CMNV2, which incorporates MobileNetV2 along with an invented CAFFE block. By simply introducing five extra layers in a pre-trained structure, this architecture is able to withstand even the most complicated situations in the real world to score 99.10% accuracy against 13 other CNN structures. The present study trained and tested the proposed CMNV2 architecture through DNNs, TL, and DL on 5,000 images that were divided into real faces and deepfake faces. When compared to the 13 CNN architectures in terms of some important parameters, the proposed model was superior in terms of accuracy, precision, recall, F1-score, error rate, and efficiency.

III. METHODOLOGY

This study presents an integrated method to detect deepfake images based on deep learning and machine learning techniques. The technique relies on the use of pre-trained CNN models to identify key features of the image, which are further used to classify them using conventional machine learning algorithms. The proposed methodology is based on a whole workflow, involving image preparation, dataset splitting, data augmentation, extraction of deep features, classification, and evaluation stages as illustrated in Fig. 1.

1) Data set collection and preprocessing: All the experiments in this study were conducted on the Deepfake Detection Dataset 2026, which was retrieved from the Kaggle portal [26] some samples images are shown in Fig. 2. It is a large dataset for detecting deepfakes, which includes facial images that are specially designed to help analyze deepfakes using machine learning methods. In the given facial image dataset, there are 6,557 images, which can be classified into two categories (2,790 real images and 3,767 fake images). The real images were obtained from Unsplash and include portraits of people of all age groups and genders. The artificial face images were generated by implementing highly sophisticated AI-based image generation algorithms, such as StyleGAN3 another. One of the key aspects that distinguishes this dataset from other commonly used deepfake image datasets is that it comes with comprehensive metadata information, which can be useful for conducting advanced analysis experiments using machine learning. The given dataset includes 17 different attributes that fall into the following categories, represented in Table I.

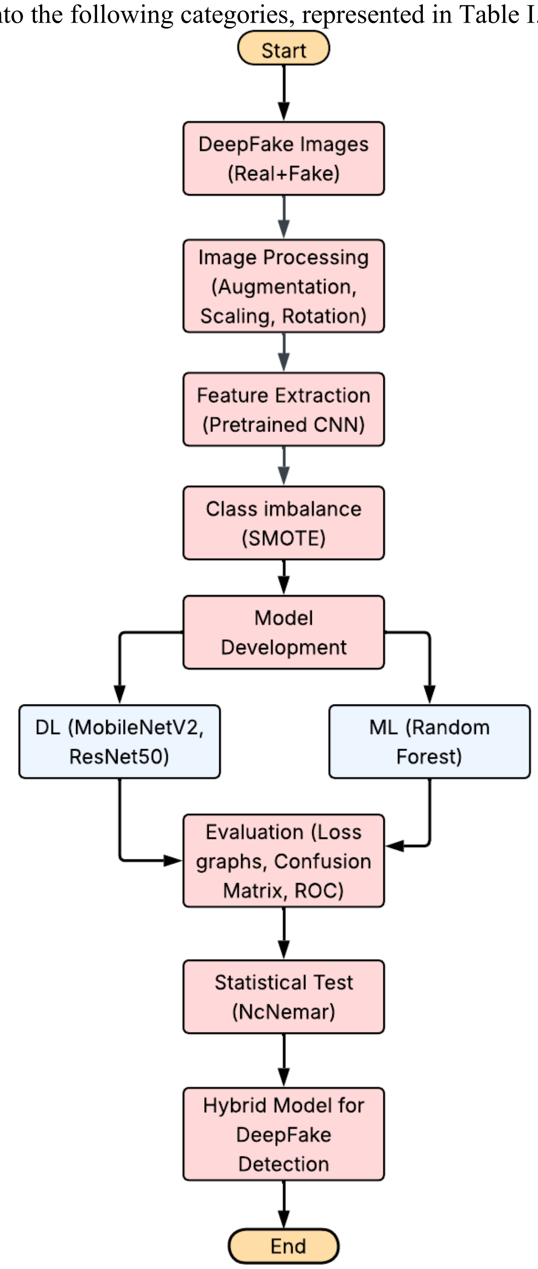

Fig. 1. Overall methodological flow chart of the proposed model.

TABLE I. DATASET CHARACTERISTICS

Dataset Characteristics	Values
Text Attributes	11
Numerical attributes	2
Unique	1
Others meta data	3
Class labels	Real or Fake
Difficulty level	Easy, Medium, Hard
Confidence score	Numeric
Source information	Text
Generation method information	

Only the high-quality faces were used in this dataset.

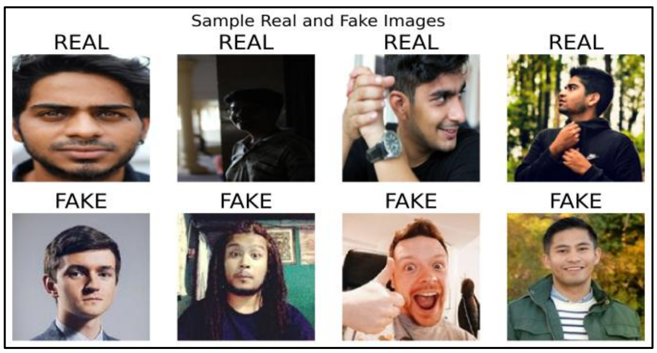

Fig. 2. Some sample images.

2) Image acquisition and preprocessing: All images are downloaded dynamically from URLs using HTTP requests. The preprocessing pipeline is shown in Fig. 3.

Fig. 3. Steps of preprocessing.

This preprocessing standardizes input dimensions and improves neural network convergence stability.

3) Data augmentation: Real-time augmentation of the dataset was done to increase the generalization capabilities of the model and minimize the risk of overfitting by applying augmentation only on the training set via the use of TensorFlow's ImageDataGenerator. Several types of augmentation were done to generate different variations of the face [27]. Fig. 4 shows the data augmentation parameters.

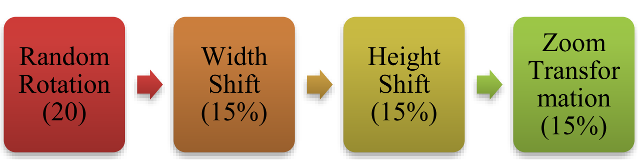

Fig. 4. Data augmentation parameters.

Fig. 4 shows the data augmentation parameters applied to images. This is to ensure that the model performs well and detects deepfake photos regardless of how the person looks in the real world.

4) Feature extraction: Feature extraction is one of the most crucial stages in the proposed system. This research utilizes pretrained convolutional neural networks (CNN) architectures as automatic feature extractors. Unlike handcrafted feature engineering, CNNs learn hierarchical spatial representations directly from facial images [28]. Given an input image be denoted as:

$$X \in \mathbb{R}^{H \times W \times C} \tag{1}$$

The conventional operation with K learnable convolution kernel and the extracted local feature map are represented as $(X * K)$, mathematically defined as:

$$(X * K)(i,j) = \sum_m \sum_n X(i-m, j-n)K(m,n) \tag{2}$$

These layers detect facial textures, blinding artifacts, compression traces, and illumination inconsistencies. The CNN extracts multilevel semantic representation using 3 layers: an early layer that learns low-level patterns using Eq. (3)-(5).

Eq. (3)-(5) represents the hierarchical feature learning conventional neural network, where each layer extracts progressively more abstract representations from the input image. The first layer captures low-level visual patterns learn more complex and discriminative features. The parameters W *and* b denote learnable weights and bias, $\sigma(.)$ denotes the nonlinear activation function.

$$\boldsymbol{F}^{(1)} = \boldsymbol{\sigma}\left(\boldsymbol{W}^{(1)} * \boldsymbol{X} + \boldsymbol{b}^{(1)}\right) \tag{3}$$

$$\boldsymbol{F}^{(2)} = \boldsymbol{\sigma}\left(\boldsymbol{W}^{(2)} * \boldsymbol{F}^{(1)} + \boldsymbol{b}^{(2)}\right) \tag{4}$$

$$\boldsymbol{F}^{(3)} = \boldsymbol{\sigma}\left(\boldsymbol{W}^{(3)} * \boldsymbol{F}^{(2)} + \boldsymbol{b}^{(3)}\right) \tag{5}$$

Furthermore, each convolution layer produces an activation map:

$$F_K = \sigma(W_K * K + b_K) \tag{6}$$

where,

- W_K: convention filters
- b_K: bias term
- σ: activation function

These feature maps are then compressed using global average pooling (GAP). For the feature map F_K :

$$g_K = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F_K(i,j) \tag{7}$$

Thereby preserving dominant semantic activation.

The final feature representation can be mathematically expressed as:

$$f = [g_1, g_2, \dots \dots . g_d] \; where \; f \in \mathbb{R}^d \tag{8}$$

This feature representation includes spatial inconsistencies, blind spots, and illumination.

The final feature vector generated through the (GAP) layer was provided as an input to the Random Forest classifier for the classification stage. The final feature vector included high-level features extracted by the network and characterized various aspects related to facial texture changes, lighting changes, compression distortions, and manipulation-related features. In order to make sure that all the features have equal weight in terms of classification, the z-score normalization was performed on the extracted features.

Dimensionality reduction techniques were not used since the GAP layer provided the compact representation of the input features. Thus, the normalized feature vectors were supplied directly into the Random Forest classifier. The classifier was designed using 100 decision trees and the Gini impurity criteria for splitting nodes. The bootstrap method was used to improve the generalization ability of the model. Moreover, the number of features for splitting was selected based on the square root rule. These configurations were chosen since they are commonly used for deep feature classification and have proved their effectiveness in the experimental analysis.

5) Handling class imbalance: There is a minor imbalance of classes, where the number of images that are not real is higher than that of the images that are real. To avoid bias towards the majority class when predicting, the class weights are assigned inversely to their frequencies, as demonstrated in Fig. 5. This was done during the training process, whereby the neural network could be trained to perform well for both classes of images. To solve the problem of unbalanced classes in the training dataset, SMOTE is employed. This approach helps synthesize more artificial samples for the class that is poorly represented by creating new samples based on existing features. Consequently, this improves the ability of the model to generalize from minority examples and eliminates any biases. To preserve the integrity of the evaluation process, SMOTE is not used for the test dataset but only for the training set to achieve the objectives [29].

6) Model development: In this study, 2 pre-trained DL models, such as MobileNetV2 and ResNet50, and 1 ML classifier, such as RF, are used.

a) MobileNetV2: MobileNetV2 was selected for this research due to its compact and efficient architecture that allows for less computational complexity. Depth-wise separable convolutions make it possible for the network to use a minimal amount of parameters and still perform efficient feature extraction.

b) ResNet50: ResNet50 was selected based on its ability to learn features in a deep network thanks to residual connections. Such an approach allows overcoming the problems associated with vanishing gradients and facilitates the process of training deep networks [30]. Models used in this research were pretrained with the ImageNet dataset with the following parameters:

- Include_top=False
- The convolutional part was not trainable during the initial training phase.

c) Random Forest Classifier (RF): RF is an ML classifier used to generate multiple decision trees. Predictions are made through a consensus of those trees, which enhances the accuracy and robustness of predictions. It is efficient in dealing with serious dimensional problems as well as in the identification of nonlinear associations between the behavioral (Fig. 6) [31].

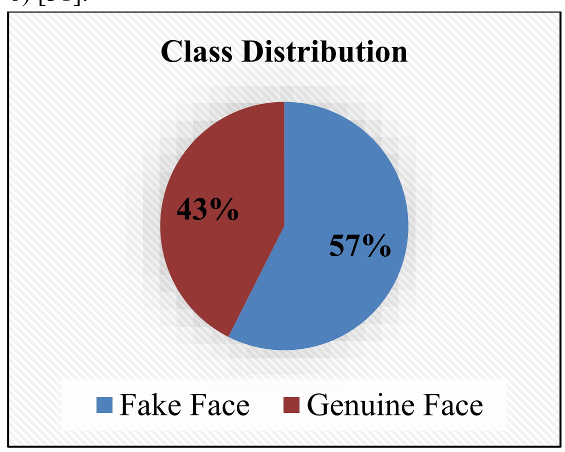

Fig. 5. Class distribution.

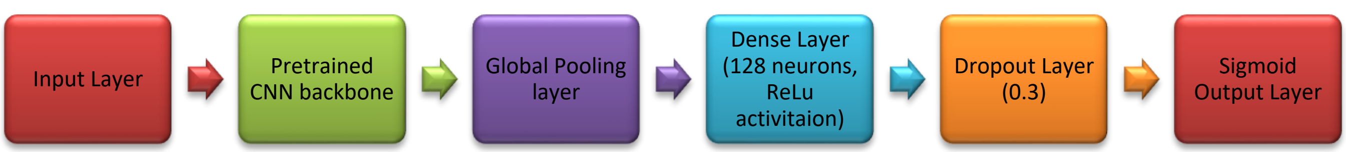

Fig. 6. Deep learning architecture parameters.

7) McNemar test: McNemar's test is a statistical procedure that can be used to make a comparison between two classifiers based on the same set of data. As opposed to testing overall accuracy, it tests the points where the two classifiers disagree. It is thus more appropriate for determining the differences between two classifiers [32].

8) Experimental setup: This experimental setup was designed to evaluate the efficacy of various ML and DL approaches that were utilized to detect real and fake faces. All experiments were conducted within the Python programming framework, utilizing typical libraries for analyzing and modeling data through machine learning techniques. Training was performed for up to 40 epochs with a batch size of 8. In order to make the training more efficient and improve the performance of the model, two callbacks were introduced.

a) Early stopping: The training was terminated automatically if no improvements in the validation loss occurred within five epochs. This ensured that training stopped at the appropriate point, avoiding any overtraining.

b) Reducing learning rate on plateau: Another callback introduced was related to the adjustment of the learning rate. If the validation loss does not change over three epochs, then the learning rate is adjusted by multiplying with 0.5 [33].

The sigmoid activation function computes the probability of fake classes by:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{9}$$

Moreover, binary cross-entropy was employed as a loss function, which can be mathematically calculated as:

$$l = -\frac{1}{N}\sum_{i=1}^{N}[y_i \; log(\hat{y}_i) + (1-y_i)\, log(1-\hat{y}_i)] \tag{10}$$

The Adam optimizer was used as a gradient optimization because of its adaptive learning-rate capability [34].

To prevent data leakage between subsets, the division of the dataset was made before the beginning of training. All the preprocessing techniques, including SMOTE oversampling and data augmentation, were applied only to the training subset. The validation and test subsets were untouched during the experiments to preserve their properties as representative data. Moreover, the feature extraction process based on the use of pretrained CNN models was performed separately on each subset. Despite the absence of identity, generation source, and near-duplicate relations labeling in the dataset, sample isolation was made during all stages of the experiment. In general, the described experimental configuration allowed obtaining unbiased results regarding the performance of the model on unseen data.

TABLE II. DISTRIBUTION OF SAMPLES

Training Samples	Validation	Testing Samples
2100	297	603

Table II depicts the distribution of samples into training and testing. To ensure consistency and reproducibility, the dataset was organized into a total samples (3000), consisting of both real and manipulated facial images. The dataset was divided into 70% (2100) samples, which were used for training, 297samples (approximately 10%) were used for validation, and 603 samples (approximately 20%) were used for testing. As for ML models, 70% used for training and 30% were used for testing, thereby maintaining consistency and enabling a comparison between the two methodologies. As far as DL methods are concerned, part of the training data set was kept for a validation data set, which helps in monitoring the progress of learning by the model, tuning the hyperparameters of the model, and minimizing the problem of overfitting. Whereas in the case of ML methodology, no separate validation data set was used because the training process involved in ML is much less complex as compared to DL methods.

IV. RESULTS AND DISCUSSION

The objective of this study is to evaluate the performance of various ML and DL classifiers for deepfake detection across multiple measuring standard metrics such as confusion matrix, loss graphs, ROC curve, etc.

TABLE III. PERFORMANCE EVALUATION OF ML MODELS

ML models	Accuracy	Precision	Recall	F1 score
MobileNetV2	98%	90%	91%	90%
MobileNetV2+RF	99%	98%	97%	97%
ResNet50	97%	93%	94%	93%
ResNet+RF	98%	94%	92%	92%

Table III presents the performance of various ML and DL models across all four measuring metrics. The obtained experimental results clearly indicated that hybrid CNNs combined with ML classifiers outperformed CNN models. The best performance was achieved by (MobileNetV2+RF) accuracy (99%), precision (98%), recall (97%), and F1score (97%). RF helps increase the performance of DL and decrease erroneous classifications by integrating feature extraction with accurate classification. Moreover, the ResNet+RF model also showed higher performance rates compared to ResNet50 alone, where accuracy was increased from 97% to 98%. This indicates that hybridization leads to improved feature extraction and discrimination. High performance of MobileNetV2+RF can be attributed to efficient feature extraction of MobileNetV2 through depthwise separable convolutions, which were followed by feature classification via RF. The results have once again proved that hybridizing CNN models with machine learning classifiers yields higher levels of stability and accurate classification compared to deep CNN-based classifiers.

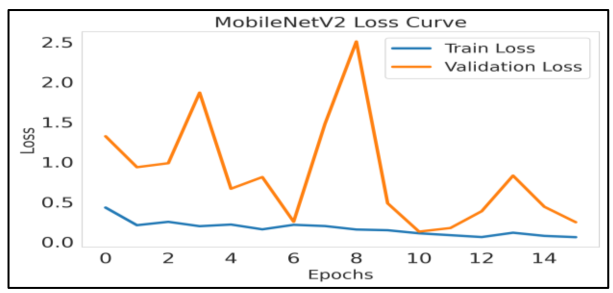

Fig. 7. Loss curves for the MobileNetV model.

Fig. 7 represents the training and validation loss of the MobileNetV2 model over 40 epochs. The decrease of MobileNetV2 training loss appears to be stable, meaning that the network successfully learns from training data and reaches good convergence during optimization. As the training loss consistently decreases, it is reasonable to believe that the network successfully extracts information from the provided training samples. The validation loss demonstrates inconsistent results and experiences sharp increases from time to time. This indicates the problem of overfitting when the neural network performs well on training data but has poor generalization capabilities when processing unknown samples. Altogether, even though the network successfully learns from training data, it requires further refinement to improve its generalization capabilities.

MobileNetV2 Validation Loss Plot demonstrates some fluctuations throughout the training process and even some spikes. Such behavior might be a hint at a minor case of overfitting, but despite that, the overall plot tends to converge, which implies that learning is taking place. Fluctuations may be associated with the dataset variation and a relatively small number of records in the validation dataset. To improve the training stability, future improvements might include more aggressive regularization, changing the dropout probability, or using data augmentation.

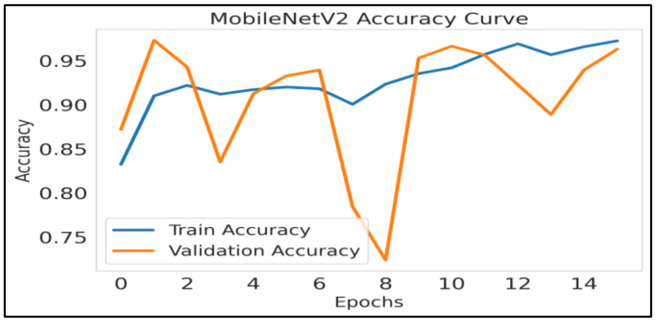

Fig. 8. Model accuracy curve.

Fig. 8 depicts the accuracy graph of the MobileNetV2 algorithm; the accuracy of the learning process by the network is good. With an increase in the epoch numbers, the training accuracy of the algorithm has been seen to increase. The validation accuracy increased along with the training accuracy; however, there have been slight differences in certain epochs due to differences in the validation data. By the last epochs of training, the validation accuracy and training accuracy are both seen to be around 96–97%. This suggests that the algorithm performed well and that there was no overfitting.

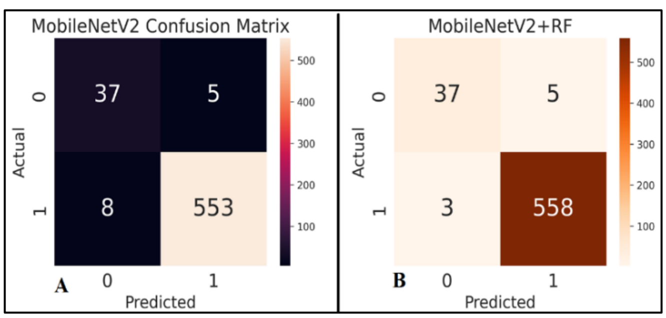

Fig. 9. Confusion matrix for MobileNetV2.

Fig. 9 shows the confusion matrix of MobileNetV2 (A) and MobileNetV2+RF Classifier (B). In Fig. 9(A), the total number of samples accurately classified as negative is 37, and the positive samples are 553. Also, 5 positive samples were wrongly classified as negative, and 8 positive samples were wrongly classified as negative. After using the RF Classifier with MobileNetV2, the correct positive samples classified were 558, and there were only 3 samples classified as negative from the positive class. There was no difference in correctly classified negative and false positive samples, which are 37 and 5, respectively. This implies that the combined model had high accuracy in terms of recall value for positive samples. The high effectiveness is due to the efficient use of the deep learning model by the RF Classifier. The hybrid model used ensemble learning to improve class separability and decrease the false positives in the positive class.

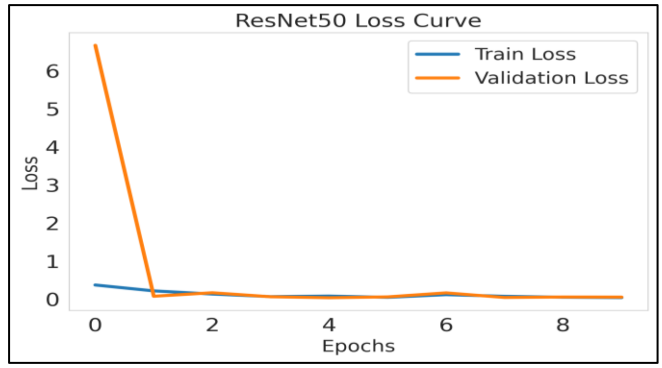

Fig. 10. Train and validation loss curves.

ResNet50 exhibits a smooth and constant learning curve throughout the entire training process, as shown in Fig. 10. Initially, the validation loss is high due to the fact that the model is trained with random weights in the early stages of training. However, following the first epoch, there is a sharp decline in the validation loss such that it is close to the training loss. The two lines are low for the rest of the training period and move almost in parallel without significant differences. Thus, the model successfully learns the critical elements of the data without overfitting. Moreover, the small gap between training and validation losses implies good generalization.

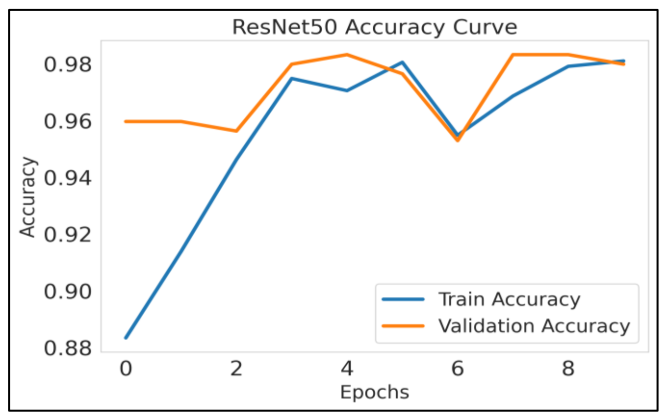

Fig. 11. Accuracy curves for ResNet50.

From Fig. 11, it can be concluded that the performance accuracy of the ResNet50 model is both consistent and accurate over 10 training epochs. The training and validation accuracy of the model increases gradually, and it eventually becomes almost 98%. Initially, the training accuracy starts at a point of 88%, and this implies that the model is learning important characteristics. The validation accuracy is fairly high and has small variations, meaning that the model does not suffer from much overfitting, and it has excellent generalization capabilities. This is evident since the training and validation graphs remain close for all epochs, meaning that the model does not experience problems related to overfitting and underfitting. There is no wide gap between the two lines, which implies that the ResNet50 model works well since it allows easy learning of deep networks. The residual learning technique makes gradient flow easy, and consequently, efficient learning becomes possible.

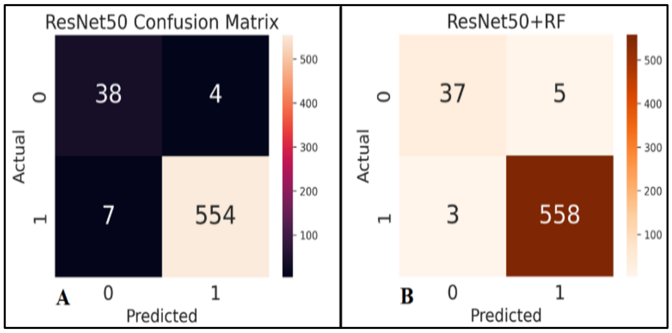

Fig. 12. Confusion matrices for ResNet50.

Fig. 12 shows the confusion matrices for standalone ResNet50 and ResNet50+RF models. Both the ResNet50 and ResNet50+RF models have an extremely high efficiency level since the majority of samples were predicted accurately. Specifically, the model of ResNet50 managed to correctly predict 38 samples of class 0 and 554 samples of class 1, misclassifying only several other samples. However, the combination of ResNet50 and RF increases the accuracy of prediction, lowering the false negatives to 3 (previously 7) and raising the number of correct predictions for class 1 from 554 to 558. It means that adding the Random Forest model to the ResNet50 algorithm significantly contributes to separating the classes correctly. Thus, the hybrid approach ensures more accurate and stable predictions than ResNet50 alone.

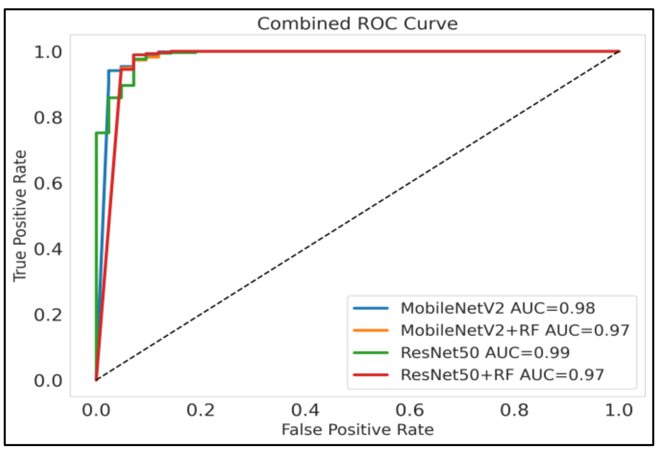

Fig. 13. ROC analysis of models.

Fig. 13 presents the ROC curve analysis results of all models analyzed. From these, the ResNet50 model managed to obtain the highest AUC value of 0.99, reflecting its superior performance in distinguishing classes. The MobileNetV2, with an AUC value of 0.98, while the RF with MobileNetV2(MobileNetV2+RF) model and the RF with ResNet50 (ResNet5+RF) model achieved an AUC value of 0.97. The findings show that the accuracy obtained by these models is quite high, and almost negligible errors were made. From these good AUC values, we can infer that our proposed deep learning models have worked well in differentiating between positive and negative classes. The ResNet50 model has performed better mainly due to its architecture, which enables efficient learning of deeper features in its structure.

TABLE IV. STATISTICAL TEST

Models	Statistic	P-Value
MobileNetV2 vs MobileNetV2+RF	3.2	0.073
MobileNetV2 vs ResNet50	0.05	0.823
MobileNetV2 vs ResNet50+RF	1.067	0.301

Table IV shows the statistical significance results of the comparative analysis among the evaluated models. The findings revealed that there is no statistically significant difference between, as all p-values are greater than the significance threshold of 0.05. Specifically, the comparison between MobileNetV2 and MobileNetV2+RF yields a test statistic of 3.2 with a p-value of 0.073, suggesting marginal improvement, this improvement is not significant. Moreover, it is possible to note that for MobileNetV2 and ResNet50, the p-value equals 0.823, while for MobileNetV2 and ResNet50+RF, it equals 0.301. In both cases, these models have almost similar classification performance levels. Thus, the analysis shows that all the methods provide rather good and quite stable results.

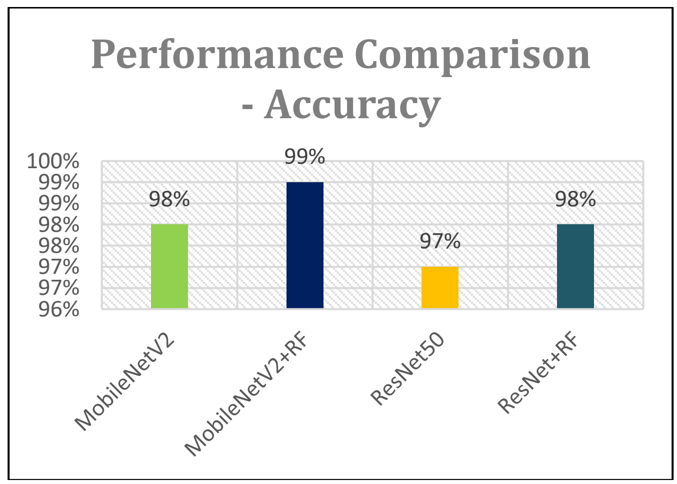

Fig. 14. Comparison of performance of ML/DL models.

In Fig. 14, the comparison of performance in terms of the accuracy metric is shown. The (MobileNetV2 + RF) achieved the highest accuracy rate (99%), attributed to its ability to classify well compared to other models. The model accuracy of MobileNetV2 and (ResNet50 + RF) was same (98%). The difference between their model accuracies is insignificant since they both differ by only 1%. Therefore, the inclusion of the RF algorithm in the deep features helps increase performance stability in classification. The accuracy of ResNet50 on its own was 97% as opposed to other models analyzed in this project. This indicates that the combination of hybrid and deep learning models provides excellent results in classification due to the high accuracy achieved.

V. DISCUSSION

The results of the experiment prove the statement that hybrid models of DL and ML provide better outcomes in terms of deepfake detection accuracy than single-layered CNN architectures, as illustrated in Table II and Fig 14. In particular, MobileNetV2 + RF appears to be the best-performing model that provides the highest accuracy rates among all other models due to the positive influence of Random Forest on CNN-based image processing. Both MobileNetV2 and ResNet50 models perform effectively without any additional machine-learning model being applied, as illustrated in Table II. The accuracy of both models equals 98% and 97%, respectively. With the help of random forest, the performance of ResNet50 increases by one percent, reaching 98%. All these data suggest the effectiveness of hybrid models, as it was mentioned above. Fig. 14 proves the effectiveness of mobileNetV2 + RF once again.

Though MobileNetV2+RF had the best numerical results compared to other models concerning most metrics, according to McNemar's paired statistical test, there is no significant difference between the performance of the models ($p > 0.05$). It simply means that the performance difference is not statistically significant. Despite the slightly higher accuracy and better stability in the results of the hybrid model, we still cannot prove that there is any practical benefit of such a network over independent convolutional neural networks statistically. Consequently, it should be understood that MobileNetV2+RF only demonstrates better trends in performance. However, considering the improvement in different evaluation criteria, we can state that hybrid CNN architectures can be used in practice to detect deepfakes.

Furthermore, ROC results illustrated in Fig. 13 confirm these findings. ResNet50 has achieved the best AUC score, equal to 0.99, whereas MobileNetV2 is a close second, having an AUC score of 0.98. Hybrid models have obtained an AUC score of 0.97. These findings suggest that all three models are good classifiers with slight differences in their ability to separate classes. Additional confusion matrix results from Fig. 9 and 12 clearly demonstrate that hybrid models effectively decrease the number of false negatives and classify images correctly even in challenging conditions, especially when the Random Forest model is applied to CNN-based feature descriptors. Stable training behavior depicted in Fig. 7, 8, 10, and 11, as well as the absence of any overfitting problems observed with ResNet50, proves that hybrid models can be considered good general classifiers.

In conclusion, it can be stated that Tables II and III, as well as Fig. 7-14 illustrate the superior performance, reliability, and robustness of hybrid models when compared to standalone CNN-based classification methods.

RQ1: How effectively do transfer learning-based CNN models like MobileNetV2 and ResNet50 help in detecting important facial characteristics to detect deepfake images?

According to the results depicted in Table II and Fig.10 and 11, both architectures, namely, MobileNetV2 and ResNet50, demonstrate excellent ability to detect discriminative facial features that could help distinguish deepfakes. The accuracy rate is high for both cases, with MobileNetV2 achieving 98% and ResNet50 reaching 97%. The learning process can be clearly traced with the help of the training and validation curves depicted in Fig. 11 (ResNet50) and Fig. 8 (MobileNetV2), where one can notice the stable and consistent behavior of the curves with a slight gap between the performance of the models during training and validation procedures. Moreover, the ROC curves plotted in Fig. 13 prove the results obtained earlier because both models show high AUC values of 0.98 and 0.99. In other words, they demonstrate their ability to differentiate real images and images that have undergone manipulations.

RQ2: Is there any improvement in the detection of deepfake images if we combine ML classifiers, such as RF, with deep learning-based facial feature extraction methods?

Indeed, the fusion of the RF classifier with feature extractors based on CNNs provides a remarkable enhancement in classification performance according to Table II and Fig. 14. The hybrid architectures, namely MobileNetV2+RF and ResNet50+RF, always surpass the performance metrics provided by the plain CNN classifiers. Specifically, MobileNetV2+RF achieves the highest value of accuracy (99%), while ResNet50+RF improves the accuracy score from 97% to 98%. One can assume that RF assists in improving the decision-making process based on the deeper feature representation obtained via the CNN architecture. Furthermore, Fig. 9 and 12 illustrate that there are fewer cases of false negatives once the Random Forest technique is applied to classification, particularly when considering the performance metrics of MobileNetV2+RF. Thus, it becomes easier to distinguish between the real and manipulated images. As a result, the proposed hybrid approach demonstrates its effectiveness since a fusion of classical machine learning and CNNs helps improve the classification process significantly.

RQ3: Which is the most suitable and accurate combination of deep learning models and ML classifiers for detecting deepfake facial images?

Based on the results obtained from the comparative analysis shown in Table II, as well as the results shown in Fig. 14 and the ROC curves in Fig. 13, it can be seen that the combination of the MobileNetV2+RF model produces the best results for deepfake video identification. This model achieves 99% accuracy, 98% precision, 97% recall, and an impressive 97% F1-score. ResNet50, on the other hand, achieves a higher area under the curve of 0.99 in Fig. 13; however, it falls short of MobileNetV2+RF in terms of performance in other evaluation parameters. The statistics provided in Table III indicate no statistically significant differences among all the models used for testing ($p > 0.05$). This means that despite some deviations, there is no strong statistical evidence regarding the differences between them. Nevertheless, MobileNetV2+RF remains the best option in terms of reliability.

Across all comparisons in this study, the p-values were consistently above 0.05. This implies that despite the small improvements observed in the hybrid models' accuracy, there is no statistical significance in the differences. Generally, the models exhibit relatively even behavior in terms of classification ability.

The proposed framework shows better results for all used metrics, but the experiments were carried out over a single dataset, which makes it difficult to be sure about how the results are generalizable to any other conditions. The absence of cross-dataset experiments and k-fold validation on various deepfake benchmarks makes it difficult to evaluate the robustness of the model to manipulation types that could be applied to deepfakes in practice. Although the presented results indicate good classification accuracy, more experiments need to be done to confirm the robustness of the model to various datasets.

1) Practical implications: The proposed approach holds practical relevance when applied to detect deepfakes. It can be applied in social networks to automatically detect manipulated images that contain fake information. The efficiency and reliability of the solution mean it could be used in cybersecurity to prevent identity theft and detect fraud. In particular, since the MobileNetV2 is rather lightweight, the solution can be applied even on devices with low computational power, like mobile phones or edge devices. Utilizing the MobileNetV2 as a feature extractor allows making the architecture of the neural network more compact and lowering the computational cost associated with running it. Additionally, the decision not to employ the full-fledged deep learning system in favor of using a Random Forest classifier significantly decreases the complexity of training the model. Additionally, due to the combination of deep learning algorithms with traditional machine learning, the approach can be scaled and used not only for images but also for other data types. The aspects like training time, inference speed, the size of the trained model, FLOPS, and memory requirements have not been evaluated. While these aspects have been considered during design, there was no formal evaluation done. Computational cost is a question that needs to be investigated in the future.

2) Limitations: Although the findings have shown high effectiveness, some limitations still exist in the experiment. First, the dataset, although diverse, is still limited in size when compared to the vast array of deepfake samples in practical situations, thus affecting the model's ability to accurately classify unknown samples that involve more complicated manipulation. Secondly, the proposed method utilizes CNN models trained by other researchers, and therefore, might not necessarily be adaptable enough for new types of deep fakes.

Although the proposed framework was successful in deepfake image recognition, the current study does not include any ablation analysis, through which the specific impact of each component would be examined. Therefore, it is difficult to understand how much the data augmentation, class balancing through the SMOTE technique, CNN-based feature extraction, and Random Forest classification have affected the outcome of the research.

Further study of these components is very important for the future. Through the process of excluding, replacing, and changing one element after another, it will be easier to see how each of the components affects the performance of the proposed framework.

Recent studies suggest that Vision Transformers (ViTs) achieve impressive results in detecting deepfakes, mainly due to the ability to encode long-range dependencies between different parts of an image. The ViT-based approaches have been mentioned in the literature review in order to represent recent advances in the field, but no comparisons with these approaches have been included in this experiment.

Conversely, the proposed study chosen to tackle the problem of detecting deepfakes is a pipeline, consisting of a pretrained CNN-based feature extractor and a Random Forest classifier. Such an architecture has lower computational complexity and is more suitable for working with small imbalanced datasets. For this reason, a comparison with ViT-based approaches has not been performed, as the latter ones have higher demands on the computational resources and a training procedure.

The future direction of this research includes the implementation of ViT-based architectures in order to compare their performance with the state-of-the-art approaches to deepfake detection.

VI. Conclusion

In this study, we presented a hybrid system that uses deep learning and machine learning for recognizing deepfake images, where deep learning with a CNN feature extractor and machine learning with Random Forest were used to improve the performance of each algorithm on its own. Experimental results revealed that MobileNetV2 + RF provided the highest value of accuracy, equal to 99%, whereas ResNet50 + RF and MobileNetV2 were among those with the best performance results. A statistical analysis of the obtained results revealed that there is no significant difference between the performances of various algorithms, implying a stable behavior of all models. Therefore, our work increases both accuracy and reliability.

Future research in this field could focus on increasing the effectiveness of transfer learning for generalizing to novel and unseen deepfake datasets that have been produced by using state-of-the-art deepfake generation techniques. It can also be interesting to explore the applicability of lightweight transformers since they can assist in learning and representation efficiency improvements. Evaluating these methods using other kinds of data, including video streams and live applications, could broaden the scope of use of the proposed system. Furthermore, efficiency improvement through model compression and pruning would enable real-time processing, especially on edge and mobile devices where computational power is limited.

Data Availability Statement: The dataset used in this study for the detection of deepfake images is publicly available at https://www.kaggle.com/code/engrgul/backdoor-ai

References

[1] P. Rathi, “Unveiling the Future: An Introduction to AI-Powered Multimedia Data Processing,” in AI Techniques for Multimedia Data Processing, IGI Global Scientific Publishing, 2025, pp. 247–268. doi: 10.4018/979-8-3693-2935-1.ch009.

[2] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, “Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning,” Information, vol. 16, no. 4, p. 270, Apr. 2025, doi: 10.3390/info16040270.

[3] C. Zhao, C. Wang, Z. Song, G. Hu, L. Wang, and D. Miao, “Multi-definition Deepfake detection via semantics reduction and cross-domain training,” Pattern Recognition, vol. 163, p. 111469, Jul. 2025, doi: 10.1016/j.patcog.2025.111469.

[4] S. Alanazi and S. Asif, “Exploring deepfake technology: creation, consequences and countermeasures,” Hum.-Intell. Syst. Integr., vol. 6, no. 1, pp. 49–60, Dec. 2024, doi: 10.1007/s42454-024-00054-8.

[5] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, “LightFakeDetect: A Lightweight Model for Deepfake

Detection in Videos That Focuses on Facial Regions," Mathematics, vol. 13, no. 19, p. 3088, Jan. 2025, doi: 10.3390/math13193088.

[6] N. Negi, D. Kaushal, S. B. Walia, and S. Fulara, "The Exposure of Deepfake Technology on Social Media and its Effects on Society," in Mastering Deepfake Technology: Strategies for Ethical Management and Security, River Publishers, 2025.

[7] I. Sagheer, "THE ROLE OF DEEPFAKE POLITICAL VIDEO IN SHAPING IDEOLOGICAL DISCOURSE ON INSTAGRAM," Journal of Applied Linguistics and TESOL (JALT), vol. 8, no. 3, pp. 356–370, Jul. 2025, doi: 10.63878/jalt968.

[8] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[9] J. Jheelan and S. Pudaruth, "Using Deep Learning to Identify Deepfakes Created Using Generative Adversarial Networks," Computers, vol. 14, no. 2, p. 60, Feb. 2025, doi: 10.3390/computers14020060.

[10] O. A. Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," Engineering Reports, vol. 7, no. 3, p. e70087, 2025, doi: 10.1002/eng2.70087.

[11] Dr. C.V. Raman University, India, A. Pandey, A. K. Tiwari, and Dr. C.V. Raman University, India, "ADVANCING FACE SPOOFING DETECTION WITH LBP, PCA, AND SVM: A ROBUST AI SECURITY APPROACH," IJIVP, vol. 15, no. 4, pp. 3589–3595, May 2025, doi: 10.21917/ijivp.2025.0508.

[12] A. M. Alashjaee, "Machine Learning Approach for Fake Image Forensics," University of Bisha Journal for Basic and Applied Sciences, vol. 1, no. 1, Sep. 2025, doi: 10.65073/3122-3508.1005.

[13] L. A. Passos et al., "A review of deep learning-based approaches for deepfake content detection," Expert Systems, vol. 41, no. 8, p. e13570, 2024, doi: 10.1111/exsy.13570.

[14] S. S. Patil, A. M, P. M, and P. P. C, "AI/ML-Based Solution for Detecting Face-Swap Deepfakes," in 2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T), Jun. 2025, pp. 1–7. doi: 10.1109/ICETI4T63625.2025.11132146.

[15] U. M and J. Bhimavarapu, "Design and development of hybrid deep learning technique for deep fake identification: in limited dataset perspective," Eng. Res. Express, vol. 8, no. 2, p. 025221, Jan. 2026, doi: 10.1088/2631-8695/ae3664.

[16] N. S. Sagheer, Z. Thamer, A. A. Abdulmunem, H. Thamer, O. Ata, and E. Mohammed, "Deepfaceguard for Real and Fake Face Classification," Apr. 2026, doi: 10.12785/ijcds/1571166529.

[17] M. S. Reza, F. Elias, M. I. Mahmud, and N. Ahmed, "Attention and residual mechanism-based CNN architecture (ARC-Net) with enhanced fairness generalization for deepfake facial image detection," PLOS ONE, vol. 21, no. 4, p. e0340099, Apr. 2026, doi: 10.1371/journal.pone.0340099.

[18] F. Castro, V. Gattulli, D. Impedovo, and A. Monaco, "DeepDect: an explainable AI platform for face swapping and face generation DeepFake detection," Multimed Tools Appl, vol. 85, no. 5, p. 500, May 2026, doi: 10.1007/s11042-026-21672-1.

[19] M. A. Arshed, S. Mumtaz, M. Ibrahim, C. Dewi, M. Tanveer, and S. Ahmed, "Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model," Computers, vol. 13, no. 1, p. 31, Jan. 2024, doi: 10.3390/computers13010031.

[20] N. A. Almalki, M. Ragab, and F. Kateb, "Leveraging AI for Sustainable Deepfake Human Face Detection Using Transfer Learning Technique," in 2025 22nd International Learning and Technology Conference (L&T), Jan. 2025, pp. 228–232. doi: 10.1109/LT64002.2025.10940123.

[21] M. M. Rahman, M. M. R. Computer, N. Das, M. S. Uddin, R. S. Shammah, and M. M. Rahman, "An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach," in 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS), May 2025, pp. 1–8. doi: 10.1109/ICBATS66542.2025.11258249.

[22] M. B. Jabra, O. Cheikhrouhou, and A. BenAmor, "Deepfake detection through ensemble learning," Multimed Tools Appl, vol. 84, no. 36, pp. 45417–45452, Nov. 2025, doi: 10.1007/s11042-025-20932-w.

[23] N. K. Alzurfi and M. S. Altaei, "A Hybrid Model of Deep Learning and Machine Learning Methods to Detect Deepfake Videos," Iraqi Journal of Science, pp. 736–750, Feb. 2025, doi: 10.24996/ijs.2025.66.2.17.

[24] A. Koçak, M. Alkan, and S. M. Arıkan, "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach," Politeknik Dergisi, vol. 28, no. 3, pp. 957–968, Jun. 2025, doi: 10.2339/politeknik.1523983.

[25] B. A. Kumar et al., "Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods," Results in Engineering, vol. 28, p. 107513, Dec. 2025, doi: 10.1016/j.rineng.2025.107513.

[26] "Deepfake Detection Challenge." Accessed: May 22, 2026. [Online]. Available: https://kaggle.com/deepfake-detection-challenge

[27] C. Fan et al., "A novel imbalance fault diagnosis method based on data augmentation and hybrid deep learning models," Structural Health Monitoring, vol. 25, no. 1, pp. 281–302, Jan. 2026, doi: 10.1177/14759217241291143.

[28] Y. Li, H.-N. Wu, H.-Z. Cheng, D.-S. Guo, and R.-P. Chen, "Application of an innovative BMask R-CNN-DAT enhanced with feature extraction and boundary awareness in tunnel leakage detection," Tunnelling and Underground Space Technology, vol. 167, p. 107074, Jan. 2026, doi: 10.1016/j.tust.2025.107074.

[29] C. Tian and R. Li, "Predicting Bond Defaults in China: A Double-Ensemble Model Leveraging SMOTE for Class Imbalance," Big Data and Cognitive Computing, vol. 10, no. 3, p. 81, Mar. 2026, doi: 10.3390/bdcc10030081.

[30] S. Riyadi, F. A. Abidin, and N. Audita, "Comparison of ResNet50V2 and MobileNetV2 Models in Building Architectural Style Classification," in 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), May 2024, pp. 1–8. doi: 10.1109/ISCV60512.2024.10620099.

[31] B. Guo et al., "Exploring the effect of the architecture morphology on urban ventilation at block scale using CFD-GIS and random forest combined method," Sustainable Cities and Society, vol. 122, p. 106241, Mar. 2025, doi: 10.1016/j.scs.2025.106241.

[32] Z. O. Aremu and J. O. Akanni, "Comparative Analysis of Mcnemar's Test and Liddell's Exact Test for Assessing Marginal Homogeneity," KolaDaisi University Journal of Applied Sciences, vol. 2, pp. 87–102, Aug. 2025, doi: 10.5281/zenodo.17438474.

[33] M. N. Razali, N. Arbaiy, P.-C. Lin, and S. Ismail, "Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets," Electronics, vol. 14, no. 4, p. 705, Jan. 2025, doi: 10.3390/electronics14040705.

[34] A. Ahmed et al., "Efficient Melanoma Detection Using Pixel Intensity-Based Masking and Intensity-Weighted Binary Cross-Entropy," International Journal of Imaging Systems and Technology, vol. 35, no. 5, p. e70179, 2025, doi: 10.1002/ima.70179.