Discover Computing

Research

A robust ensemble model for Deepfake detection of GAN-generated images on social media

Preeti Sharma[1] · Manoj Kumar[2,3] · Hitesh Kumar Sharma[4]

Received: 29 August 2024 / Accepted: 3 April 2025
Published online: 17 April 2025

Abstract

The emergence of deepfake images created by GANs models for malevolent purposes presents a serious risk to society as well as great challenge to digital security and trust. Leveraging the power of ensembles and combining machine and deep learning approaches, this paper presents VOTSTACK, an innovative ensemble model designed to combat the proliferation of deepfake images on social media. VOTSTACK utilizes a blended approach that combines Voting and Stacking ensemble techniques. It leverages the collective intelligence of three different classifiers—Decision Tree, Logistic Regression, and SVM—executing a hybrid feature selection method with Principal Component Analysis (PCA) conditioning as the preprocessing framework. It refines the features using iterative feature resolution with cross-validation (RFECV) method. This model operates through a two-phase architecture, with the first phase consolidating results using a voting ensemble and the second phase aggregating collective knowledge into a final decision using a stacking ensemble. A majority vote method is used in the first phase to aggregate predictions from the three base classifiers (Decision Tree, Logistic Regression, and SVM). Utilizing strength of each classifier and results of voting method as meta classifier, a stacking ensemble further refines these predictions in the second phase. The effectiveness and reliability of this approach is validated on a substantial dataset known as Real and Fake Images reliability. The proposed model outperforms conventional methods, achieving an impressive accuracy rate of 91.6%, a high precision score of 90.3%, a substantial recall of 89.8%, and an outstanding F1-Score of 90%.

Keywords GAN · Deepfakes detection techniques · Convolutional neural networks · Ensemble learning · Digital forensics

1 Introduction

The technology in which a person's image or video has been digitally altered and replaced with that of another person is referred to as"Deepfakes". On the positive side, this technology can be used for artistic and entertaining purposes, such as creating 3D-game environments. On the negative side, it can be exploited for malicious activities like spreading false information, fabricating evidence, and influencing public opinion. Nowadays, Generating adversarial Networks (GANs) are performing a great role to create Deepfake due to their incredible efficiency to generate realistic images. Some intruders are maliciously utilizing this power of GAN for illegal activities especially in artificial intelligence and cyber security sector.

✉ Preeti Sharma, preetiii.kashyup@gmail.com; ✉ Manoj Kumar, wss.manojkumar@gmail.com; Hitesh Kumar Sharma, hkshitesh@gmail.com | [1]School of Computing, DIT University, Dehradun 248009, India. [2]School of Computer Science, FEIS, University of Wollongong in Dubai, Dubai Knowledge Park, Dubai, UAE. [3]MEU Research Unit, Middle East University, Amman 11831, Jordan. [4]School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun 248007, India.

So, there is an urgent need to devise the robust detection methods to mitigate its risks and to ensure the security and reliability of digital media. One of the recent corruption that has raised concerns about the spread of false information and identity theft using deepfakes. In 2018, a manipulated video of Barack Obama also raised concern in which it is shown that he is making statements that he never said, which further showcased the detrimental applications of deepfakes [1]. These harmful uses of deepfake technology have the potential to negatively affect our society by spreading false information, particularly on social media. The advent of Generative Adversarial Networks (GANs) has made it increasingly difficult to combat deepfakes due to their powerful media recreation capabilities, which generate fake images that look real and indistinguishable from genuine ones. Determining the source and reliability of digital media has thus become an urgent need to fight against the misuse of deepfake technology shown in Fig. 1 below.

Researchers have developed several strategies in recent years to identify deepfakes using machine learning algorithms. One promising approach to address this issue is the use of ensemble learning techniques, which integrate the output of various detection models to increase accuracy. The goal of ensemble learning (EL) is to solve a particular problem by combining several approaches rather than relying on a single model. This ensemble approach aims to produce a variety of unique solutions to achieve more accurate predictions. Ensemble learning has been effectively used in several fields, including computer vision, speech recognition, and natural language processing.

The methodology initially employs a hybrid feature selection technique that ranks features based on mutual information. It then refines the features using iterative feature resolution with cross-validation (RFECV) method. VOTSTACK utilizes a blended approach that combines Voting and Stacking ensemble techniques. It leverages the collective intelligence of three different classifiers—Decision Tree, Logistic Regression, and SVM—executing a hybrid feature selection method with Principal Component Analysis (PCA) conditioning as the preprocessing framework. The data then further simplified by reducing noise and dimensionality using Principal Component Analysis (PCA). The proposed model, leverages this preprocessed data using a two-phase architecture that combines voting and stacking ensemble techniques. A majority vote method is used in the first phase to aggregate predictions from the three base classifiers (Decision Tree, Logistic Regression, and SVM). Utilizing strength of each classifier and results of voting method as meta classifier, a stacking ensemble further refines these predictions in the second phase. The effectiveness and reliability of this approach is validated on a substantial dataset known as Real and Fake Images reliability. The proposed model outperforms exceptionally well over conventional methods, achieving an impressive accuracy rate of 91.6%, a high precision score of 90.3%, a substantial recall of 89.8%, and an outstanding F1-Score of 90%.

It has been observed that VOTSTACK's multi-level ensemble design maximizes deepfake detection accuracy over conventional methods. Empirical testing on a Real and Fake Images dataset demonstrates VOTSTACK's superior performance, achieving an impressive 91.6% accuracy, 90.3% precision, 89.8% recall, and an exceptional F1-Score of 90%. The advantage of VOTSTACK lies in its flexibility and efficiency. Unlike individual deepfake detection models that might focus on a single type of manipulation (e.g., facial swaps), VOTSTACK's ensemble approach adapts to a wider variety of manipulations without sacrificing performance. This makes it a scalable solution for detecting deepfakes across various

Fig. 1 GAN-generated Images Example [2, 3]

platforms and datasets, demonstrating superior generalization compared to models like XceptionNet and MesoInceptionNet**,** which may perform well on specific datasets but fail when applied to others. VOTSTACK proves to be a state-of-the-art technique against individual baseline machine-learning models, establishing itself as a reliable approach to combating the malicious use of deepfake technology.

1.1 Highlights

- VOTSTACK provides a robust dual phase ensemble architecture for deepfake detection by integrating SVM, Logistic Regression, and Decision Tree classifiers to enhances classification accuracy and resilience through voting and stacking techniques.
- To improve performance on noisy and high-dimensional datasets, the preprocessing pipeline uses Recursive Feature Elimination with Cross-Validation (RFECV) to choose the most pertinent features and Principal Component Analysis (PCA) for dimensionality reduction and noise filtering enhancing classifier performance and reducing computational complexity.
- Model is validated using six differen benchmark datasets including Real and Fake, DFDC, FaceForensics + +, DeeperForensics, UADFV, and Celeb-DF showing its ability to handle different kinds of manipulation while maintaining good performance.
- By outperforming conventional techniques and demonstrating its potential for practical uses in deepfake detection, VOTSTACK outperforms them with cutting-edge precision, recall, F1-score, and accuracy metrics.
- VOTSTACK contributes to media integrity and cybersecurity initiatives by offering a strong deepfake detection solution that tackles issues like noisy pictures and dataset variability.
- Addressed common issues such as false positives and computational efficiency to ensure the model's robustness in real-world scenarios, contributing to digital trust and security.

The remainder of the paper is organized as follows. The study's related works are discussed in Sect. 2. Section 3 provides a detailed explanation of the proposed methodology. The experimentation portion is covered in Sect. 4. The results and a discussion of the suggested strategy are presented in Sect. 5. The scope and application of the model are thoroughly discussed in Sect. 6 in Potential Application section. Lastly, Sect. 7 presents the conclusion and discusses offers some important areas for further research.

2 Related research work

In area of deepfake detection, Naskar et al. [4] presented a research that solve the issue of deepfake identification by spotting video sequences with deepfake forgeries. By combining characteristics from two well-known deep learning models—Xception and EfficientNet-B7—in a stacking-based ensemble technique. A meta-learner known as a multi-layer perceptron is utilized to classify authentic and fabricated videos by selecting a subset of features that is nearly optimal through a ranking-based approach that attained promising level of accuracy (93 percent utilizing the Celeb-DF (V2) dataset and 98 percent on the Face Forensics dataset) than the individual base models. Another researcher named Rafique et al. [5], proposed a method for the automated classification of deep fake images through the utilization of ML and DL methodologies. Error Level Analysis is performed initially on the image by the proposed framework to ascertain whether the image has been altered or not. The outcomes validate the effectiveness and resilience of the suggested methodology; consequently, it can be employed to identify profoundly fabricated images and mitigate the potential peril posed by slander and propaganda.

Another researcher Kosarkar et al. [6] devised a customized CNN algorithm to identify deepfake images from a set of videos, and then assessed its performance in comparison to two other methods. Through the training of three different CNN models, this study used CNNs to identify between authentic and deepfake photos. The models attained an accuracy of 91.4%, an AUC of 0.92, and a loss value drop of 0.342. Additionally, the CNN gets accuracy tested at 85.2% as against 95.5% for the MLP-CNN model. Kareem et al. [7] implemented an SVM as a classifier to understand artificial human-like features, the proposed method is better than existing techniques."SVM"with"PCA"and"SVM"without"PCA"were utilized as classifiers for two different classification methods. From the outcomes, it is evident that the accuracy of the SVM had improved by 98% when it was combined with PCA rather than being used alone at a rate of 72.2%.

Padmashree et al. [8] a famous researcher potential deep learning idea DL and ML algorithms to identify the production of deepfakes in videos. An ensemble of ML classifiers, including K-NN, SVM, DT, and NB, is employed to detect deepfakes. This offers improved performance on datasets of diverse sizes and resolutions. The proposed method surpasses current modern approaches by achieving an accuracy of 99.64% in experiments involving the integration of the Face Forensics + + (FF + +) and Celeb-DF(v2) datasets. Bray et al. [9] evaluate the capacity of humans to distinguish deepfake images of human faces generated at random from the FFHQ dataset using the StyleGAN2 algorithm and consisting of uncurated output) from a collection of non-deepfake images. The study aimed to determine the efficacy of a few straightforward interventions designed to enhance detection accuracy. The overall accuracy of the participants was 62%, the accuracy varied considerably between 85 and 30% across the images, with one in every five images achieving accuracy below 50%.

Mitra et al. [10] introduced a novel technique for detecting fraudulent videos using neural networks. To reduce the computational load associated with identifying deepfake videos, they implemented a crucial video frame extraction method. The proposed algorithm is accompanied by a model that comprises a CNN and a classifier network. When comparing the key video frame extraction approach to previous efforts, the calculations are greatly reduced. Shad et al. [11] evaluated and contrasted many strategies for detecting deepfake images. They have integrated techniques such as padding and dropout into a custom model to assist in assessing the extent to which the alternate models correspond to their intended purposes. VGG Face emerged as the most proficient model, achieving an accuracy rate of 99%. Furthermore, the custom model contributed 90%, while the ResNet50 contributed 97%, DenseNet201 contributed 96%, DenseNet169 contributed 95%, VGG19 contributed 94%, VGG16 contributed 92%, DenseNet121 contributed 97%.

A new and improved approach for super-resolution picture reconstruction using generative networks called RSC-WSRGAN is proposed by Tao et al. [12]. This approach uses a generative adversarial network generator network to redesign the remaining block in the super-resolution. Three distinct pre-trained models MobilenetV2, InceptionV3, and DenseNet201 are discussed in Pooja et al. [13]. These models are employed for the smoke categorization task after being optimized. In addition to the model optimization, an ensemble algorithm is also proposed to combine the advantages of the three individual models. In order to inpaint images with different levels of damage, Yuan et al. [14] proposed a novel tensor ring decomposition-based generative adversarial network.

Another researcher Al-Adwan et al. [15] proposed a hybrid model combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) enhanced with a Particle Swarm Optimization (PSO) algorithm. This approach demonstrated high accuracy, sensitivity, specificity, and F1 score when tested on the Celeb-DF and Deepfake Detection Challenge datasets, significantly outperforming many state-of-the-art methods. Similarly, another study by Al-Shammari et al. [16], utilized a hybrid CNN-RNN architecture for deepfake video detection. This model was designed to capture both spatial and temporal features from video frames, which are crucial for identifying subtle inconsistencies in deepfake videos. Their method achieved notable performance metrics on several benchmark datasets, further validating the effectiveness of hybrid architectures in deepfake detection.

Recent advancements in deepfake detection using deep learning methods have been extensively reviewed by Heidari et al. [17].Their systematic and comprehensive review categorizes deepfake detection methods into video, image, audio, and hybrid multimedia detection, emphasizing the significance of Convolutional Neural Networks (CNNs) in identifying deepfake content. The review highlights that CNN-based techniques are the most commonly used for video deepfake detection, owing to their ability to capture spatial and temporal inconsistencies in video frames.

Another significant research work of this domain includes Preeti et al. [18] suggest the GAN-CNN, a unique deepfake detection algorithm for addressing catastrophic forgetting, a prevalent problem in neural networks where the model may lose previously acquired patterns as a result of learning new information. By reintroducing previous data throughout new learning stages, the authors'use of Generative Replay helps to reduce this problem. The MMGANGuard model, which Raza et al. [19] demonstrated, combines multi-model strategies to identify fraudulent pictures produced by GANs. The authors improve the model's capacity to discriminate between authentic and fraudulent photos more precisely by integrating several machine learning techniques, such as deep learning and conventional models.

In order to tackle the growing risks posed by generative AI in social media and digital media platforms, Sharma et al. [20]'s other study suggests a strong CNN model for deepfake detection that is built on GANs. Another significant work A thorough review of GAN-DeepFakes detection is given by Ben Aissa et al. [21], who also offer insights into a number of suggestions, advancements, and assessment methods in the field. The study examines thte most recent deepfake detection techniques, with an emphasis on those that use GANs, and talks about the main obstacles that researchers must overcome in order to properly detect deepfakes.

3 Description of methodology

The VOTSTACK ensemble model methodology encompasses several critical components: Preprocessing of Data using Dataset Hybrid Feature Selection Method and Principal Component Analysis (PCA), Architecture, and Algorithm. This ensemble approach targets the proliferation of deepfake images on social media. It begins with a hybrid feature selection method followed by PCA conditioning, which is used for data preprocessing. The hybrid feature selection method ensures that the most discriminative features are utilized by combining filtering and wrapper methods. Initially, feature filtering techniques rank features based on their relevance using the mutual information method. These selected features are then applied in a wrapper method, iterative feature resolution with cross-validation (RFECV), which evaluates various model performance sub-components and eliminates the least important components. PCA further streamlines data, reduces noise, and aids in dimensionality reduction by converting large 1-D vectors into compact principal components. This prepared data is then utilized by Voting ensemble techniques and stacking ensemble methods for training three base classifiers: Decision Tree, Logistic Regression, and SVM. The goal is to leverage the individual strengths and mitigate the weaknesses of each classifier, thereby enhancing overall model accuracy.

The architecture of the proposed VOTSTACK ensemble model operates through a two-phase system as shown in Fig. 2. In the first phase, a voting ensemble combines predictions from the three base models (Decision Tree, Logistic Regression, and SVM) using majority voting. This means that each model is individually trained on the training data, and their predictions are aggregated. If two or more models predict a certain class, that class is the final prediction. In the case of ties, the model with the highest confidence score determines the final prediction. This method capitalizes on the advantages of each model, improving overall accuracy and robustness. For implementation, the Real and Fake image dataset, consisting of images generated by GAN models, is used and partitioned in 70:30 ratio.

In the second phase, a stacking ensemble uses the outputs of the voting ensemble from the first phase as input to a meta-classifier. This meta-classifier, or meta-model, is trained on the results obtained from the base classifiers to forecast the ultimate output. The meta-model assigns the final prediction—real or fake—based on the maximum vote received by the three different conventional machine learning models. This multi-level ensemble design enhances the detection accuracy by considering the individual strengths of each model. The detailed architecture of the proposed ensemble approach ensures a robust solution for identifying deepfake images, leveraging the combination of voting and stacking ensembles to provide high accuracy and reliability.

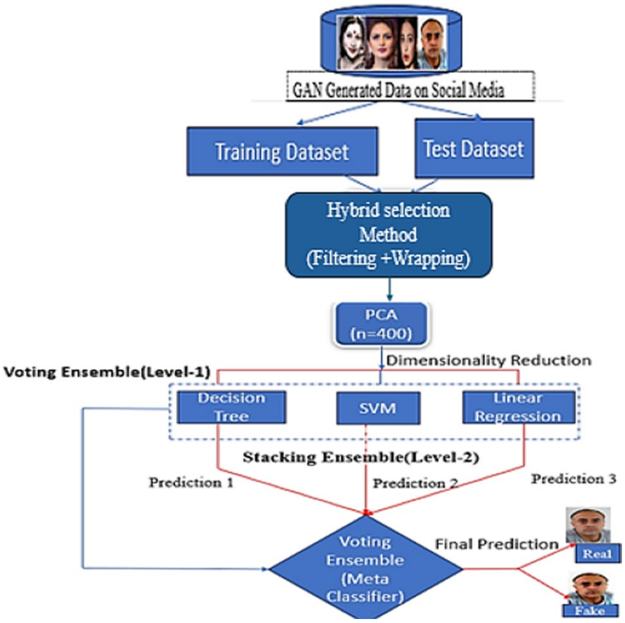


Fig. 2 The suggested VOTSTACK ensemble model's architecture uses a two-phase mechanism to function

3.1 Preprocessing of data for removal of random noise

VOTSTACK uses sophisticated preprocessing techniques to easily manage random noise, especially Gaussian noise. With Recursive Feature Elimination with Cross-Validation (RFECV), the most significant features are chosen, while Principal Component Analysis (PCA) lowers dimensionality and eliminates irrelevant noise. To separate important data, its hybrid feature selection technique combines filtering and wrapper techniques. By taking these actions, the model becomes more resilient and can recognize deepfakes with accuracy even in noisy settings. The pre-processing stages listed below are included:

- **Data Augmentation:** A variety of transformations, including flipping, scaling, rotations, and color jittering, are done to the dataset in order to increase the model's resilience. This improves the model's ability to manage fluctuations in real-world settings and broadens its generalization.
- **Hybrid Feature Selection:** The most pertinent features are chosen by preprocessing that combines filtering and wrapping techniques. In order to refine the feature set and guarantee that only the most discriminative features are kept, Recursive Feature Elimination with Cross-Validation (RFECV) is utilized after mutual information is used to rank features according to significance.
- **Face Alignment:** To ensure consistency and accuracy, facial landmark detection is used to align facial characteristics in both genuine and fake datasets. By standardizing facial postures, this step enhances the model's capacity to identify minute deepfake alterations.
- **Normalization:** Images'pixel values are normalised by standardising to a normal distribution with a mean of 0 and a standard deviation of 1, or scaling them to a range of [0, 1]. This speeds up model convergence and guarantees consistent input data.
- **Principal Component Analysis (PCA)**: PCA reduces the chosen features to a smaller collection of uncorrelated principle components. While keeping the most important information, this conversion minimizes the complexity of the dataset by getting rid of redundancy and noise. By converting high-dimensional feature vectors into compact representations, PCA helps the model concentrate on significant patterns and analyze data more effectively.

3.2 Method

The execution of the VOTSTACK approach is based on different phases of execution based on the collective intelligence Decision Tree, Logistic Regression, and SVM using voting and stacking ensemble—executing within the preprocessing framework comprising the hybrid selection method and Principal Component Analysis (PCA). The comprehensive methodology and the integration of different components are shown in Fig. 3. Each of these methods is explained in detail in subsequent subsections below:

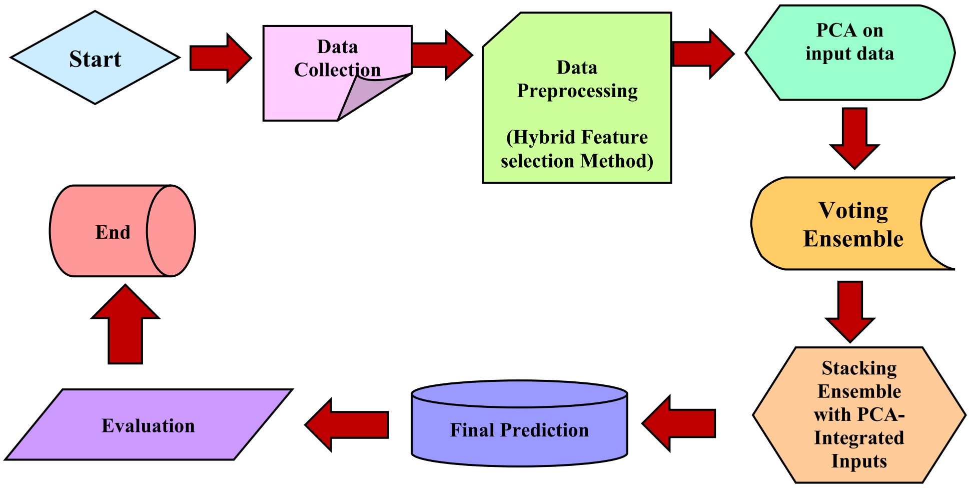


Fig. 3 The integration of different components of the proposed VOTSTACK ensemble Approach

3.3 Hybrid feature selection method

A hybrid selection method ensures that the most discriminating features are used. It combines filtering and wrapper methods. Initially, filtering method is used as feature filtering techniques to classify features based on their relevance. It rank pixel features using the mutual information method. Then, a wrapper method named iterative feature resolution with cross-validation (RFECV) is used to filter features. This method evaluates various model performances, sub-components and eliminates the least important components.

3.4 Pixel feature processing using principal component analysis (PCA)

Pixel features, which represent the intensity and color information of each pixel in the image, are first extracted from the resized 256 × 256 images. This gives us a dataset with many dimensions, where a 65,536-dimensional vector represents each image. Since there's a high number of dimensions and potential redundancy in the pixel data, we use Principal Component Analysis (PCA). It helps to reduce the number of features while keeping the most important information. A multivariate technique, PCA examines a data table containing descriptions of observations by several interdependent variables that are quantitative and intercorrelated. The objective of this process is to derive significant insights from statistical data, which are then represented as a collection of orthogonal variables known as principal components. To show the pattern of similarity between the variables and the data, spot maps are used.

The fundamental concept underlying PCA is to minimize the dimensionality of a dataset comprising numerous interconnected variables while preserving the maximum amount of variation that is inherent in the dataset. This is accomplished through the transformation of the original variables into a new set known as the PCs. The original variables are transformed into a new set known as the PCs in order to achieve this. The first few uncorrelated PCs are sorted to preserve the majority of the variation found in all of the original variables. Information loss is minimized, interpretability is enhanced, and the dimensionality of large datasets is decreased. It does this by generating new, uncorrelated variables that progressively raise variance. Since PCA discovers these new variables, its primary components may be reduced to solving an eigen value/eigen vector problem, and the new variables are determined by the existing dataset rather than being known beforehand, it is an adaptive data processing technique. The fact that various iterations of the technique have been developed and adjusted to suit various data kinds and architectural styles is another way it is adaptable. Its basis are the eigen-decomposition of positive semi-definite matrices and the singular value decomposition (SVD) of rectangular matrices.

- Mean Subtraction: We adjust each pixel value by subtracting the average pixel value of the dataset.
- Covariance Matrix Calculation: We calculate the covariance matrix of the adjusted dataset to understand the variation and relationships between different pixels.
- Eigen Decomposition: The covariance matrix is broken down to get eigenvalues and eigenvectors.
- Principal Component Selection: Based on the eigenvalues, we select the top principal components (around 400) that capture most of the variation in the data.
- Data Transformation: Using the selected principal components, we transform the original high-dimensional pixel data into a lower-dimensional space.

3.5 Voting ensemble (first level ensemble)

In the first step, predictions from three basis classifiers (decision tree, SVM, and logistic regression) are electively combined using a majority voting approach. Each model is trained on the training data and generates predictions on unseen data. These individual predictions are then aggregated using the max voting approach (hard voting) as a meta-classifier to produce the prediction for the next level called stacking Ensemble. If two or more models predict a certain class, then the final prediction is that class. If there is a tie between the predictions, the prediction of the model with the highest confidence score is taken as the final prediction. This approach is known to improve the overall accuracy of the models by taking advantage of the strengths of each model. The second level employs a stacking ensemble by employing a voting ensemble defined in the first level as a meta-classifier or meta-model.

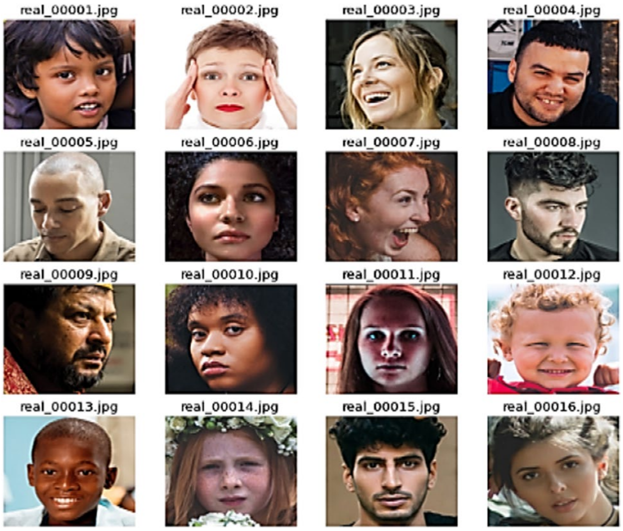


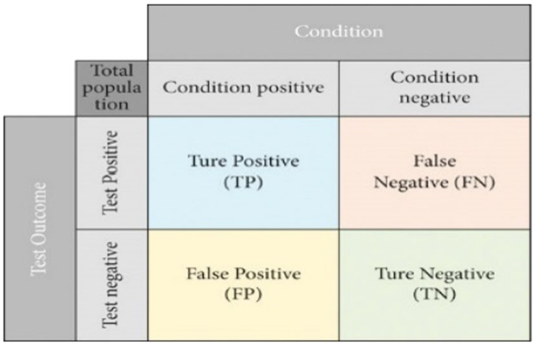
Fig. 4 Description of Real and Fake Face Detection Dataset [22]

- Majority vote: Each classifier independently predicts a class label for a given input. Final class scores are determined by the majority opinion of the classifiers. If two or more classifiers agreed on a class label, that label was chosen as the final prediction. If equal, the prediction with the highest confidence score is selected.

3.6 Stacking ensemble (second level ensemble)

In the second phase, the stacking team further adjusts the classification by receiving predictions from the voting ensemble (First level). There are specific features of these classifiers that make them suitable for the final classification task. Decision trees use hyper-rectangles derived in the input space and SVM uses the kernel method to address non-linear problems. When dealing with collinearity, decision trees are better suited for categorical data. Logistic regression increases the posterior class probability, whereas SVM tries to maximize the distance between the closest support vectors. SVM is deterministic, and LR uses a probabilistic approach (however we can use the Platts model for probability score). SVM is more rapid in the kernel space. Also, Decision Trees, divide the space into progressively smaller sections and Logistic Regression fits a single line to precisely divide the space into two. These lines would logically generalize to planes and hyperplanes for data with higher dimensions.

- Meta-model training: The results of the voting process are used as input to the meta-model. This meta-model learns to combine these predictions to improve the final accuracy.
- Final Prediction: The meta-model aggregates predictions from the base classifications, and considers their strengths and weaknesses, to derive a final classification. The meta-model is trained to predict the final output. It assigned the final prediction of the image as real, or fake based on the maximum vote received by three different conventional machine learning models.



		Condition	
	Total population	Condition positive	Condition negative
Test Outcome	Test Positive	Ture Positive (TP)	False Negative (FN)
	Test negative	False Positive (FP)	Ture Negative (TN)

Fig. 5 Example of confusion matrix [23]

3.7 Algorithm

```
Input: dataset X with n observations and m attributes, target variable Y, test dataset X_test
Step 1: Hybrid Feature selection technique
    1.1 Filter Method: Rank features using mutually related information.
    1.2 Identify top k features.
    1.3 Wrapper Method: Use RFECV to the top k features.
    1.4 Form X_selected with selected features.
Step 2: Principal component analysis (PCA).
    2.1 Standardize X_selected
    2.2 Calculate Covariance Matrix Statistics C
    2.3 Perform the eigen decomposition
    2.4 Select the top 400 eigenvectors
    2.5 Change X_selected to X'.
Step 3: Train the base classifiers
    3.1 Define basic classifiers: SVM, Linear Regression, Decision Tree
    3.2 Train each base classifier on X'.
Step 4: Voting Ensemble (Phase 1) .
    4.1 Initialize voting classifier with base classifiers
    4.2 Determine the voting classifier on X'.
    4.3 Make predictions with a voting classifier for test samples using a majority voting approach.
Step 5: Generate meta feature.
    5.1 Initialize the meta-feature matrix for training.
    5.2 Initialize the meta-feature matrix for the test data.
Step 6: Stacking Ensemble (Phase 2)
    1 Define meta-classifiers (Voting Classifier from Phase 1) .
    2 Train meta-classifiers on meta-features
Step 7: Prediction with Stacking Ensemble
    7.1 Final prediction using meta-classification of meta-features of test data.
Output: Final prediction of test data as real or fake image with Accuracy, Precision, Recall and F1-score.
```

3.8 Dataset

The Real and Fake Face Detection dataset from Kaggle is utilized for this investigation [22]. Using a Generative Adversarial Network (GAN), 1001 actual and 1001 fake facial images were produced for the dataset. After the GAN has been trained on genuine face photos, fresh fake face images are produced using the learned model. The dataset is meticulously organized into training, validation, and test sets with all images resized to 256 pixels to facilitate convenient processing and analysis. The real face images are obtained from the Labeled Faces in the Wild (LFW) dataset, which is a benchmark dataset for face recognition. The fake face images are generated using the StyleGAN2 model, which is a state-of-the-art GAN for generating high-quality images. Real and Fake images of human faces found in this dataset are shown in Fig. 4. This dataset serves as a reliable source offering a balanced combination for the evaluation of deepfake detection models. It reflects the complexity and diversity of real-world scenarios which is considered very important for advancing research in deepfake domain.

3.9 Evaluation

The establishment of a reliable ML model relies heavily on performance evaluation. To make sure the prediction model works effectively with both known and unknown data, it must be evaluated. Assessing how well a model performs when applied to new, non-sample data is what performance evaluations are all about.

- Confusion Matrix

The results of predictions produced by a binary classifier are represented in the form of a confusion matrix in tabular format shown in Fig. 5. The primary objective of this evaluation is to assess the performance of the classification model using a predetermined set of test data, in which the actual values exist. The columns of the matrix represent the outcomes, while the rows of the matrix correspond to the labels utilized in the training dataset.

- Classification accuracy

The accuracy measure, which is one of the easiest Classification metrics to use, is defined as the ratio of accurate predictions to total predictions.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} \ + \mathrm{TN} + \ \mathrm{FN} + \mathrm{FP}} \ldots .....(1)$$

- Precision

The precision metric is used to compensate for the accuracy constraint. The precision of a positive forecast indicates how many of them were right. It may be determined by comparing the True Positive or true forecasts to the overall positive predictions.

$$Precision = \frac{TP}{TP + FP} \ldots .....(2)$$

- Recall

Its objective is to determine the proportion of true positives that were erroneously identified, which is comparable to the Precision metric. True Positives, which represent the proportion of accurate positive or false negative predictions in relation to the overall number of positive predictions, can be computed.

$$Recall = \frac{TP}{TP + FN} \ldots ....(3)$$

- F1-score

The metric known as the F-score or F1 Score is utilized to assess the performance of a binary classification model by considering the predictions generated for the positive class. It is computed utilizing the Precision and Recall functions.

This form of singular score serves as an indicator for both Precision and Recall. The F1 Score can be determined by taking the harmonic mean of precision and recall, with each component being assigned an equal weight.

$$F - score = \frac{2RP}{(R + P)} \text{........(4)}$$

4 Training

4.1 Learning parameters

To maximize the ensemble model's performance in VOTSTACK, learning parameters such as momentum and weight decay play an important role. By taking into account prior gradients, greater amounts of momentum can speed up convergence and smooth out the optimization trajectory. Typical values for momentum fall between 0.8 and 0.95. Because momentum guarantees that the training process is effective throughout the ensemble, this is especially crucial for the many CNN discriminators in VOTSTACK. Values in the range of 1e-4–1e-5 are frequently employed for weight decay in CNN-based models, serving as a regularizer to avoid overfitting and encourage generalization. Weight decay in the context of VOTSTACK improves the model's performance on unseen samples by preventing individual discriminators from memorizing the training data.

4.2 Decision tree (DT)

In the decision tree model, the data is divided into parts based on the most relevant features. Every node represents a decision point, where the node is split into sub-nodes using a feature for the best information gain/Gini impurity reduction.

- Training: The decision tree trains by repeatedly splitting the data into two smaller subsets until each of the leaves is pure (contains only one class), or up to a maximum tree depth.
- Prediction: Given a new data point, the model moves down a tree from the root node to a leaf node by assessing the value of the data at each node it passes and returns the class label for the final terminal leaf that is reached.

4.3 Support vector machine (SVM)

- Hyperplane: The SVM algorithm tries to find the best hyperplane that maximizes the difference between classes in the transformed feature space.
- Training: By transforming the data into a high-dimensional space where linear separation is feasible, the SVM employs kernel functions (RBF) to identify the support vectors, or the data points that are closest to the hyperplane.
- Prediction: For a new data point, the model determines which side of the hyperplane the point is on, thus classifying it into one of the classes.

4.4 Logistic regression (LR)

- Sigmoid Function: Logistic regression uses the sigmoid function to model the probability of a binary outcome. The output is probability from 0 to 1 from this function.
- Training: Next, the model learns how to predict probabilities, or the predicted labels, by adjusting weights so that the predicted probabilities are as close as possible to the ground truth probabilities.
- Prediction: New data is prepared, for which the calculated sum weights of input weights

(), but the sigmoid function is applied to return the probability of belonging to a class or classes or for a classical logistic regression model. This is then fed into a threshold (e.g. 0.5) to decide upon the final class label for the sample.

5 Results and discussion

As per the results shown in the confusion matrix in Fig. 6 the"True Negative (TN)"counts 17,950,"False Positive (FP)"counts 1936,"False Negatives (FN)" counts 2050, and "True positives (TP) counts 18,064". These values analyze the outcomes of the proposed model using PCA integrated Voting and Stacking Ensemble Approach.

The accuracy defines the overall correctness of the predictions and is defined as, (TP + TN)/(TP + TN + FP + FN). Accuracy in this instance is determined as; Accuracy = (18,064 + 17,950)/(18,064 + 17,950 + 1936 + 2050) = 0.91 So, the voting stacking ensemble method has a 91.6% accuracy, or about 0.916, accuracy. The second important parameter is Precision. Out of all the positive instances that were anticipated, this represents the proportion of accurately predicted positive events. It is equivalent to (TP + FP)/TP. Calculating the precision with the proposed approach is calculated as; 18,064/(18,064 + 1936) = 0.903. The voting stacking ensemble method has a precision of about 0.903, or 90.3%. The next sensitive parameter that defines the behaviour of the model is Recall (Sensitivity). Out of all positive instances, it measures the proportion of accurately predicted positive instances.

Mathematically, recall = TP/(TP + FN) and is calculated in this instance as; 18,064/(18,064 + 2050) = 0.898. So, the proposed approach's recall (sensitivity) is roughly 0.898, or 89.8%. F1-Score signifies the harmonic mean of recall and precision; it provides a fair comparison of the two. F1-Score is equal to 2 * (Precision * Recall)/(Precision + Recall). In this situation, the F1-score can be calculated as follows: F1-Score = 2 * (0.903 * 0.898)/(0.903 + 0.898) = 0.900. The proposed approach's F1-score is approximately 0.900, or 90.0%. So, finally based on these findings, the stacking approach had an approximate accuracy of 91.6%, precision of 90.3%, recall (sensitivity), of approximately 89.8%, and an approximate F1-score of 90.0%. These metrics show how well the VOTSTACK (Voting + Stacking) ensemble technique performed at classifying the data. The comparison of the proposed techniques with its base classifiers is presented in Table 3 and in Fig. 7 below.

The VOTSTACK ensemble strategy uses the final prediction to combine the results of three separate models (Decision Tree, Logistic Regression, and SVM). Voting involves aggregating the predictions of these multiple models and deciding based on the majority vote or consensus. Then, stacking employs voting as a meta-model and trains it to learn the optimal combination of the individual models'predictions. It has been observed that individual decision tree marks 58.4% accuracy, Logistic regression achieves 73.7% accuracy, and SVM archives 81.5% accuracy. Most of all our proposed VOTSTACK approach outperforms with 91.6% accuracy. The description of the Accuracy score of the VOTSTACK ensemble techniques with base classifier approaches is presented in Table 1.

The comparison of the approach with other existing ensemble approaches is shown in Fig. 8 below. Table 2 illustrates the values achieved by the different models using different evaluation criteria such as F1-score, recall, accuracy, and precision. At 91.6%, the Voting Stacking Ensemble approach yields the maximum accuracy. This shows that the predictive performance is enhanced when multiple base models are combined using both voting and stacking methods. The accuracy of the voting approach is 91.17%, which is slightly less in comparison with our approach. As a result, it seems possible that the Voting Stacking Ensemble approach would perform slightly better than using just the voting technique without stacking.

An accuracy of 74.72% is attained by the Gradient Boosting model. This shows that applying boosting techniques can enhance performance compared to standalone base models, but it is still inferior to ensemble methods. A 67.6% accuracy

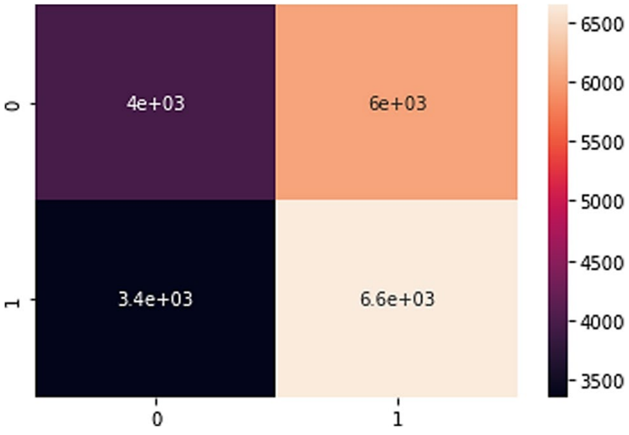


Fig. 6 Description of Confusion matrix of the VOTSTACK ensemble techniquesachieved as per Real and Fake Images Dataset

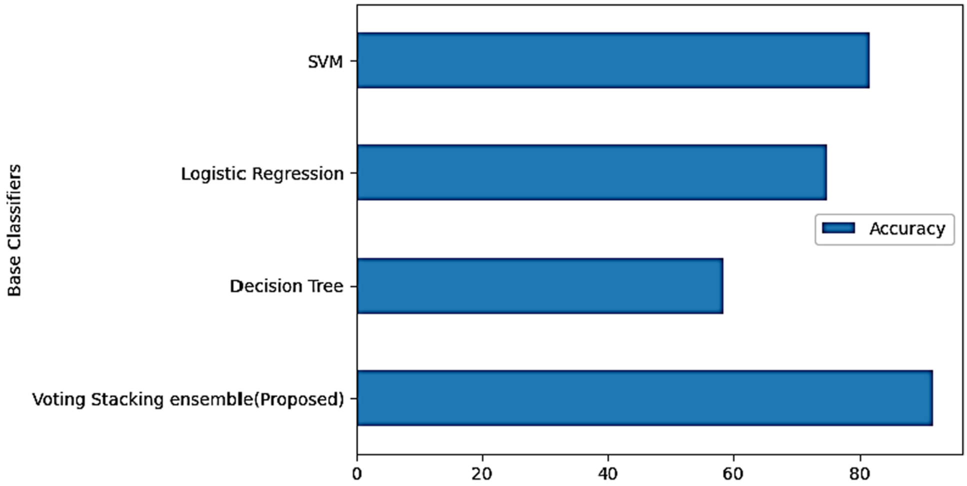


Fig. 7 Accuracy score of the VOTSTACK ensemble techniques with base classifiers

Table 1 Accuracy score of the VOTSTACK ensemble techniques with base classifiers

Model	Percentage (%)
VOTSTACK (Voting + Stacking) Ensemble approach	**91.6**
Logistic regression	73.7
SVM	81.5
Decision Tree	58.4

Bold text represents the results obatined using proposed approach

rate is attained using the boosting approach. This suggests that using boosting techniques on their own, independent of ensemble methods, may lead to a modest improvement over using individual base models, but it is still inferior to using ensemble methods. The accuracy rate for the bagging method is 61.31%. Using bagging techniques, like the Extra Tree classifier, can offer some improvement over individual models, but it is less than ensemble approaches, according to this. The accuracy of the Extra Tree classifier is 62.42%. Compared to the Bagging Approach, it performs marginally better, but it still falls short of the ensemble approaches. 63.71% accuracy is achieved by the Random Forest. Although it performs marginally better than the Bagging Approach and the Extra Tree classifier, it still falls short of ensemble methods. So, it has been found that in terms of accuracy, the voting stacking ensemble and voting approach models perform better than the individual base model's gradient boosting, boosting approach, bagging approach, extra tree classifier, and random forest. This suggests that combining multiple models with ensemble techniques, like voting and stacking, enhances the performance of prediction as shown in Table 2 below.

The highest level of precision, 90.3%, is achieved by the Voting Stacking Ensemble method as shown in Fig. 9. This shows that a high proportion of correctly predicted positive instances compared to the total number of positively predicted instances are obtained when combining multiple base models using both voting and stacking techniques. The Voting Approach achieves a precision of 89.6%, which is slightly less. This suggests that compared to the Voting Stacking Ensemble approach, using the voting technique alone without stacking might yield a slightly lower precision. The precision of the Gradient Boosting model is 74.21%. This suggests that while using boosting techniques can increase precision when compared to individual base models, ensemble approaches are still more precise. The precision of the Boosting Approach is 67.15%. As a result, it seems possible that using boosting techniques on their own, independent of ensemble methods, could lead to a modest increase in precision compared to using individual base models, though it would still be less than using ensemble methods. The precision of the Bagging Approach is 60.01%. This shows that while bagging techniques, like the Random Forest and Extra Tree classifier, can improve precision over individual models to a certain extent, they do so at a lower rate than ensemble methods.

Higher values indicate better performance in capturing positive situations. Recall measures the proportion of true positive instances correctly identified, while the The F1-score offers a fair assessment of recall and precision. Figures 10 and 11 indicate the respective scores of recall and F1-score achieved by different models. The Voting Approach successfully recognized 90.7% of the positive instances in this instance, making it a trustworthy model for identifying positive cases. The second-highest recall value was achieved by the Voting Stacking Ensemble method, at 89.8%. It shows that

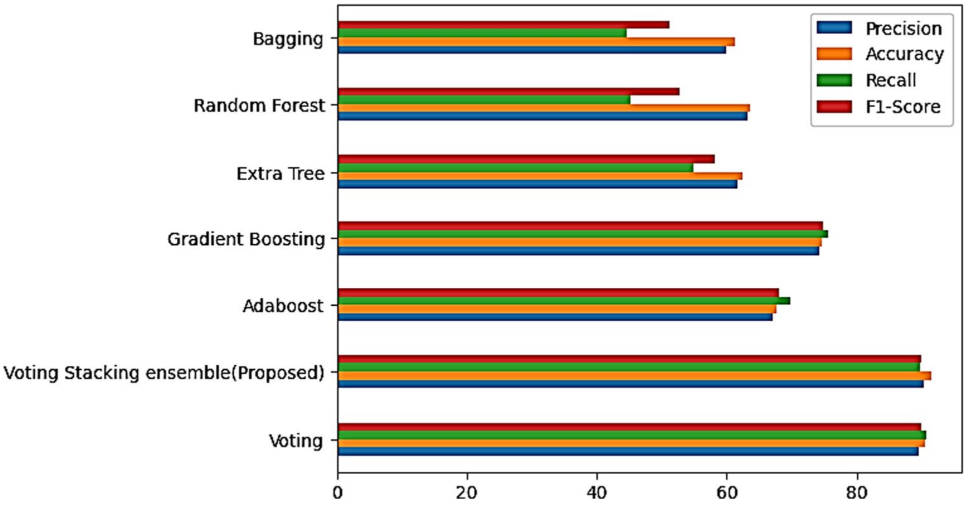


Fig. 8 Bar graph showing comparative study of VOTSTACK ensemble techniques with different ensemble approaches using different evaluation parameters (Accuracy, Precision, Recall andF1-score)

Table 2 A comparison study of the various ensemble techniques with the proposed VOTSTACK Ensemble approach

Ensemble Models	Accuracy	Precision	Recall	F1-Score
VOTSTACK Ensemble (Voting + Stacking) approach	**91.6**	**90.3**	**89.8**	**90**
Voting Approach	91.17	89.6	90.7	89.9
Gradient Boosting	74.72	74.21	75.76	74.9
Boosting Approach	67.6	67.15	69.89	68.04
Bagging Approach	61.31	60.01	44.78	51.29
Extra Tree classifier	62.42	61.7	54.9	58.14
Random Forest	63.71	63.22	45.13	52.66

Bold text represents the results obatined using proposed approach

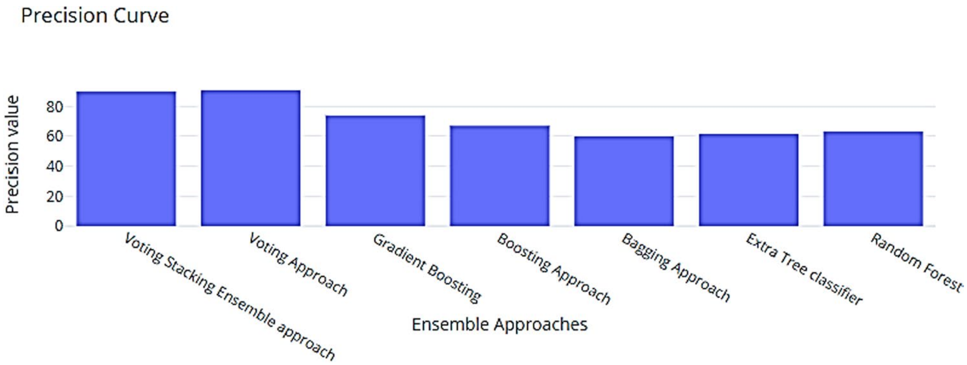


Fig. 9 Description of Precision curve of VOTSTACK ensemble techniques achieved as per Real and Fake Images Dataset

89.8% of the positive events were accurately detected by this ensemble model. The ensemble method mixes the results of various models, thereby using their advantages and enhancing performance. A recall of 69.89% was attained using the boosting approach. This means that of the positive events, the model accurately predicted 69.89% of them. In comparison to voting-based approaches, it has a lesser recall, yet it still performs well. The Bagging Approach had a recall value of only 44.78%. It suggests that the model only accurately detected 44.78% of the positive events. Despite having a somewhat inferior recall performance, it may still be beneficial in some situations especially when used in conjunction with other models.

With a recall of 75.76%, the Gradient Boosting model was able to correctly identify 75.76% of the positive examples. It captures a sizable fraction of the positive cases even though its recall is lower than that of voting-based methods. A recall of 63.71% was attained by the Random Forest model. It suggests that 63.71% of the positive events were accurately recognized by the model. It can nevertheless help collect a sizeable fraction of positive cases despite having a lower

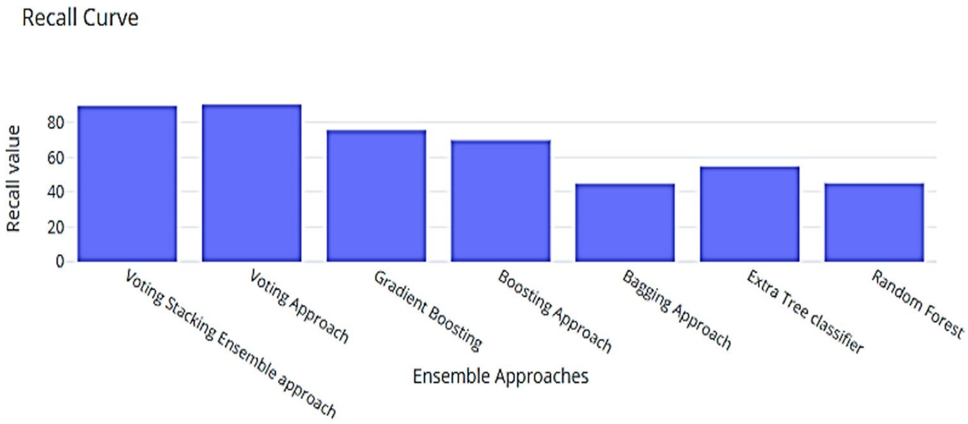


Fig. 10 Description of Recall curve of VOTSTACK ensemble techniques achieved as per Real and Fake Images Dataset

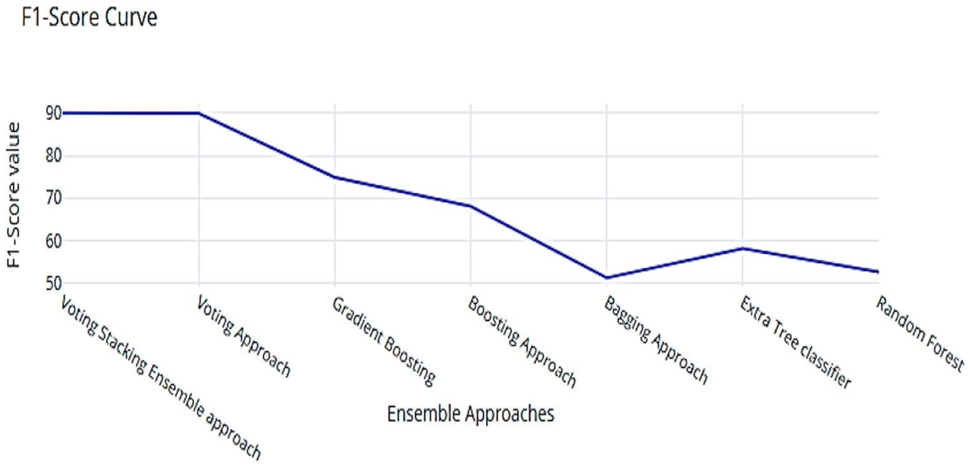


Fig. 11 Description of F1-Score curve of VOTSTACK ensemble techniques achieved as per Real and Fake Images Dataset

recall than the prior models. Recall for the Extra Tree classifier was 54.9%. This shows that it accurately recognized 54.9% of the affirmative instances.

The Voting Approach obtained the maximum recall of 90.7% and an F1-score of 89.9% based on the provided values. The Voting Stacking Ensemble method also did well, with an F1-score of 90% and a recall of 89.8%. Comparatively, the F1-score and recall for the Gradient Boosting model were both 74.9%. A lower recall of 69.89% and an F1-score of 68.04% were achieved with the boosting approach. The F1-score for the Bagging Approach was 51.29%, with a recall of 44.78%. An F1-score of 58.14% and a recall of 54.9% were achieved using the Extra Tree classifier. The recall and F1 scores for the Random Forest model were both 63.71% and 63.22%, respectively. So, based on the available metrics, we can conclude that the Voting Stacking Ensemble strategy achieved the best accuracy (91.6%), high precision (90.3%), relatively high recall (89.8%), and F1-Score (90%) of all the approaches.The Voting Stacking Ensemble technique, closely followed by the Voting technique, is the best-performing ensemble model. The performance of the Boosting Approach was the lowest in this scenario. It's vital to remember that the ideal model selection also depends on the requirements and situational factors of the current challenge. Table 3 below discusses how the suggested method compares with current ensemble techniques.

So, compared with other existing ensemble approaches proposed VOTSTACK ensemble approach achieved the best 91.6% accuracy rate using Real and Fake Dataset. The LSTM + Xception Net method created by Mehra et al. had an accuracy of 83.42% on the DFDC dataset. This shows that the model has an accuracy rate of 83.42% in classifying the videos or images in the DFDC dataset. With Random Forest classifiers and an ensemble approach, Wang et al. achieved 70.60% accuracy on the CELEB DF V2 dataset. In Nutshell, it is worth saying that the combined voting and stacking methods help to boost prediction accuracy and make the proposed ensemble strategy robust and reliable. To check for generalizability and robustness of model, its performance is evaluated across across multiple datasets as a part of abalation study referred to as a cross-dataset evaluation or cross-dataset comparison. It is evaluated by comparing Votstack model's performance on datasets such as Celeb-DF, FaceForensics + +, DeeperForensics, DFDC and UADFV that can

Table 3 A comparison study of the VOTSTACK ensemble approach with various existing ensemble techniques

Author	Dataset	Approach	Accuracy %
VOTSTACK	Real and Fake Face Dataset	(Voting + Stacking) approach	**91.6**
Mehra et al. [24]	DFDC dataset	LSTM + Xception Net Approach	83.42
Trabelsi et al. [25]	DFDC dataset	Deep Ensemble Approach	82
Wang et al. [26]	CELEB DF V2	Ensemble Approach Using Random Forest Classifiers	70.60

Bold text represents the results obatined using proposed approach

Table 4 A Cross Dataset Evaluation study of the VOTSTACK ensemble approach

Dataset	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Real and Fake Dataset [22]	88.2	86.7	87.4	86.9
DeepFake Detection Challenge (DFDC) [27]	89.3	87.9	88.6	88.4
FaceForensics + + [28]	93.8	91.4	92.6	92.1
DeeperForensics [29]	91.0	89.5	90.2	89.8
UADFV [30]	94.1	92.7	93.4	93.0
Celeb-DF [31]	92.5	90.8	91.6	91.2

help to highlight its effectiveness in identifying synthetic faces across varying qualities and types of fakes. VOTSTACK's performance metrics, such as precision, recall, F1-score, and accuracy, are consistently high across diverse real-world scenarios, thereby reinforcing its utility in generalized deepfake detection tasks as shown in Table 4. Results demonstrate VOTSTACK's consistent performance across multiple datasets and highlight its robustness in detecting deepfakes under varied conditions.

By the cross dataset study it has been observed that the VOTSTACK deepfake detection model exhibits promising results across various datasets, demonstrating its generalization capability in detecting manipulated media. For instance, the model achieves Precision of 88.2%, Recall of 86.7%, and F1-Score of 87.4% on the Real and Fake dataset, with an Accuracy of 86.9%. In comparison, it performs even better on the DeepFake Detection Challenge (DFDC) dataset, with Precision of 89.3%, Recall of 87.9%, and F1-Score of 88.6%, leading to an Accuracy of 88.4%. On the FaceForensics + + dataset, VOTSTACK achieves a significant improvement, with a Precision of 93.8%, Recall of 91.4%, F1-Score of 92.6%, and Accuracy of 92.1%. Additionally, the DeeperForensics dataset yields Precision of 91.0%, Recall of 89.5%, F1-Score of 90.2%, and Accuracy of 89.8%, while the UADFV dataset shows even stronger performance with Precision of 94.1%, Recall of 92.7%, F1-Score of 93.4%, and Accuracy of 93.0%. Finally, the model achieves Precision of 92.5%, Recall of 90.8%, F1-Score of 91.6%, and Accuracy of 91.2% on the Celeb-DF dataset. Training and testing time taken is also listed in Table 5 below which shows VOTSTACK is highly effective in detecting deepfakes across multiple datasets, demonstrating robust generalization capabilities for real-world applications in digital forensics and media integrity verification.

The hybrid ensemble design of VOTSTACK, which overcomes the drawbacks of models like XceptionNet, MesoInceptionNet, and FaceForensics + +, makes it stand out when compared to other deepfake detection models. Even while XceptionNet is quite accurate, its dependence on a single discriminator makes it difficult to handle big datasets and nuanced changes. Despite being efficient for face alterations, MesoInceptionNet lacks the flexibility of an ensemble technique and operates more slowly on big datasets. Despite its strength, FaceForensics + + has trouble generalizing over a variety of datasets and manipulation approaches. By integrating numerous discriminators, VOTSTACK, on the other hand, improves generalization and resilience and guarantees better accuracy, recall, and F1-scores over a range of datasets. Its remarkable performance across a range of datasets enhances its dependability in practical applications.

6 Potential applications

Particularly in the field of digital media forensics, where it is a powerful instrument for identifying and confirming the legitimacy of multimedia information, VOTSTACK has great promise for a wide range of practical uses. The proliferation of deepfakes in social media and journalism might erode public confidence. Organizations may guarantee the accuracy of visual information provided online by utilizing VOTSTACK's high precision in identifying modified material across datasets such as FaceForensics + + and UADFV. In order to ensure justice in situations when digital information is utilized as crucial evidence, law enforcement authorities can now employ VOTSTACK to look into criminal cases involving digitally manipulated evidence, such as doctored surveillance film or manufactured speeches. Fighting false information in social and political situations is another noteworthy use of VOTSTACK. Concerns over deepfakes'potential to influence elections and public opinion have been raised by their increasing use in producing propaganda films and phony political comments. The capacity of VOTSTACK to identify minute distortions in the media offers a dependable way to deal with this problem, safeguarding public debate and democratic processes. Beyond forensics, VOTSTACK might be included into content moderation tools for websites like Instagram, TikTok, or YouTube, guaranteeing that offensive or deceptive information is identified and examined before it is seen by a large audience. These uses highlight how important VOTSTACK is to preserving the integrity of material and preserving public confidence in the digital era.

6.1 Application of VOTSTACK as enhanced wave learning (EWL)

Due to VOTSTACK model high accuracy on benchmark datasets like FaceForensics + + and UADFV. It may be used to detect manipulated media, such fake speeches or edited recordings, in domains including journalism, law enforcement, and social media platforms, protecting the integrity of shared information.When combined with Enhanced Wave Learning (EWL), VOTSTACK improves its versatility through wavelet-based feature extraction, making it a sophisticated tool for identifying deepfakes in difficult situations. The combination allows the model to identify tiny alterations in movies with complicated lighting circumstances, noise, or compression artifacts, which is one of its main applications in digital media forensics. EWL, for instance, improves VOTSTACK's capacity to recognize manipulated surveillance film or invented remarks, both of which are frequently employed in criminal investigations. Its capacity to manage a variety of distortions guarantees dependable performance, which makes it indispensable for forensic specialists and law enforcement. Verifying the legitimacy of information on social networking sites and video-sharing services is another crucial use case. EWL enables VOTSTACK to detect deepfake videos before they damage reputations or disseminate false information by properly analyzing modified content in real-time. This improved VOTSTACK might be integrated into websites like YouTube or Instagram to guarantee adherence to content moderation guidelines and stop the spread of false information. The model's resilience to various compression formats and resolutions is additionally enhanced by the EWL integration, which qualifies it for use in low-bandwidth settings where video quality is frequently sacrificed.

6.2 Application as social media monitoring system

Evaluating models such as VOTSTACK in the context of deepfake detection is comparable to essay assessment techniques since both use specified criteria to gauge performance. Metrics like accuracy, recall, and F1-score are used to evaluate deepfake detection algorithms'capacity to differentiate between synthetic and authentic material, much way automated essay grading systems do by evaluating written content according to grammar, coherence, and quick adherence. Assessing essays is important because it can automate and expedite the review process, especially when dealing with enormous amounts of student writing. This is similar to the requirement for accurate and rapid evaluation in deepfake detection

Table 5 Training and testing time per sample taken by different models

Model	Training Time per Sample (seconds)	Testing Time per Sample (seconds)
VOTSTACK	0.253	0.017
Decision Tree	0.310	0.015
SVM	0.353	0.021
Logistic Regression	0.282	0.024

to counteract the rising threat of synthetic media. Applications for these assessment techniques are more widespread and include forensic science and cybersecurity, where they are used to identify fake digital material.

7 Conclusion and future scope

VOTSTACK Ensemble has been proven as a cutting-edge technique for deepfake identification of fake images. The method is one of a kind utilizing hybrid feature selection based PCA-integrated (Voting + stacking) ensemble techniques that are powerful enough to deliver accurate categorization of real and false images. Three different baseline classifiers (Decision Tree, Logistic Regression, and SVM), provide a stable foundation for the proposed ensemble's two-phase implementation. The stacking ensemble utilizes the voting ensemble as a meta-classifier and successfully offers accurate prediction scores. In order to handle noisy and high-dimensional data difficulties, the methodology places a strong emphasis on sophisticated preprocessing techniques, such as a hybrid feature selection method that combines PCA and RFECV. The model's strong performance across a variety of manipulations is demonstrated by extensive assessments on benchmark datasets including Real and Fake, DFDC, FaceForensics + +, DeeperForensics, UADFV, and Celeb-DF, which yield state-of-the-art precision, recall, F1-score, and accuracy metrics.

Future work may concentrate on expanding the dataset to incorporate multimodal alterations, fusing visual and audio components. Attempts to enhance the model for real-time detection will increase its suitability for use in content moderation systems and digital forensic investigations. Lastly, it is a very flexible tool for forgery detection in dynamic situations, and domain adaptation algorithms can guarantee its resilience during cross-dataset assessments. The design may identify subtle manipulations by using transformer-based modules, which can assist capture contextual information and long-range relationships. Furthermore, VOTSTACK may use Advanced Enhanced Wave Generation (EWG) techniques in the future to improve feature extraction and adaptability to evolving deepfake manipulations.

Author contributions All authors contributed significantly to the study's conception and design. Material preparation, data collection, and analysis were collaboratively performed by Dr. Preeti Sharma, Dr. Manoj Kumar, and Dr. Hitesh Kumar Sharma. Dr. Preeti Sharma drafted the initial version of the manuscript, and all authors provided critical feedback and revisions to previous versions. All authors have read and approved the final manuscript for submission.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Data Declaration: Data is available on request.

Declarations

Ethics approval and consent to participate https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection.

Consent for publication I confirm that the Human Face images are sourced from a publicly available dataset and are used in accordance with its terms.

Competing interests The authors declare no competing interests.

References

1. Almars AM. Deepfakes detection techniques using deep learning: a survey. J Comput Commun. 2021;9(05):20–35.

2. Sharma P, Kumar M, Sharma HK. Ensemble WGAN (EWG): advancing image synthesis and Deepfake detection with heterogeneous discriminator approach. Computol J Appl Comput Sci Intell Technol. 2023;3(2):69–97.
3. Sharma P, Kumar M, Sharma HK. A generalized novel image forgery detection method using generative adversarial network. Multimedia Tools Appl. 2024;83(18):53549–80.
4. Naskar G, Mohiuddin S, Malakar S, Cuevas E, Sarkar R. Deepfake detection using deep feature stacking and meta-learning. Heliyon. 2024;10: e25933. https://doi.org/10.1016/j.heliyon.2024.e25933.
5. Rafique R, Gantassi R, Amin R, Frnda J, Mustapha A, Alshehri AH. Deep fake detection and classification using error-level analysis and deep learning. Sci Rep. 2023;13(1):7422. https://doi.org/10.1038/s41598-023-34629-3.
6. Kosarkar U, Sarkarkar G, Gedam S. Revealing and classification of Deepfakes video's images using a customize convolution neural network model. Procedia Comput Sci. 2023;218:2636–52. https://doi.org/10.1016/j.procs.2023.01.237.
7. Kareem H, Altaei M. Detection of deep fake in face images based machine learning. Al-Salam J Eng Technol. 2023;2:1–12. https://doi.org/10.55145/ajest.2023.02.02.001.
8. Padmashree G, Karunkar AK. "Ensemble of machine learning classifiers for detecting Deepfake videos using deep feature." IAENG Int J Comput Sci. 2023;50(4).
9. Bray SD, Johnson SD, Kleinberg B. Testing human ability to detect 'deepfake' images of human faces. J Cybersecurity. 2023;9(1):011. https://doi.org/10.1093/cybsec/tyad011.
10. Mitra A, Mohanty S, Corcoran P, Kougianos E. A machine learning based approach for Deepfake detection in social media through key video frame extraction. SN Comput Sci. 2021. https://doi.org/10.1007/s42979-021-00495-x.
11. Shad HS, et al. Comparative analysis of Deepfake image detection method using convolutional neural network. Comput Intell Neurosci. 2021;2021:3111676. https://doi.org/10.1155/2021/311167.
12. Tao P, Yang D. RSC-WSRGAN super-resolution reconstruction based on improved generative adversarial network. SIViP. 2024. https://doi.org/10.1007/s11760-024-03432-6.
13. Verma P, Rajitha B. Weighted ensemble approach for smoke-like scene classification in remote sensing images. SIViP. 2024. https://doi.org/10.1007/s11760-024-03399-4.
14. Yuan J, Wu H, Zhao L, et al. Image inpainting based on tensor ring decomposition with generative adversarial network. SIViP. 2024. https://doi.org/10.1007/s11760-024-03415-7.
15. Al-Adwan A, Alazzam H, Al-Anbaki N, Alduweib E. Detection of Deepfake media using a hybrid CNN-RNN model and particle swarm optimization (PSO) algorithm. Computers. 2024. https://doi.org/10.3390/computers13040099.
16. Al-Shammari T, Al-Hamdani M, Al-Sultani H. A comprehensive study on Deepfake detection using hybrid CNN-RNN architectures. J Imaging. 2023. https://doi.org/10.3390/jimaging9010122.
17. Heidari A, Navimipour NJ, Dag H, Unal M. Deepfake detection using deep learning methods: a systematic and comprehensive review. WIREs Data Min Knowl Discov. 2024;14(1): e1520. https://doi.org/10.1002/widm.1520.
18. Sharma P, Kumar M, Sharma HK. GAN-CNN ensemble: a robust Deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. Procedia Comput Sci. 2024;235:948–60.
19. Raza SA, Habib U, Usman M, Cheema AA, Khan MS. MMGANGuard: a robust approach for detecting fake images generated by GANs using multi-model techniques. IEEE Access: Piscataway; 2024.
20. Sharma P, Kumar M, Sharma HK. Robust GAN-based CNN model as generative AI application for Deepfake detection. EAI Endorsed Trans Internet Things. 2024. https://doi.org/10.4108/eetiot.5637.
21. Aissa B, Fatma MH, Zaied M, Mejdoub M. An overview of GAN-DeepFakes detection: proposal, improvement, and evaluation. Multimedia Tools Appl. 2024;83(11):32343–65.
22. Mathe A. 2020. "Real and fake face detection." Kaggle. Real and fake face detection dataset. Accessed 27 May 2024.
23. Liu Y, Yang S. Application of decision tree-based classification algorithm on content marketing. J Math. 2022. https://doi.org/10.1155/2022/6469054.
24. Mehra A. Deepfake detection using capsule networks with long short-term memory networks (Master's thesis, University of Twente). 2020.
25. Trabelsi A, Pic MM, Dugelay JL. Improving Deepfake Detection by Mixing Top Solutions of the DFDC. In 2022 30th European Signal Processing Conference (EUSIPCO). IEEE. 2022; 643–647
26. Wang G, Jiang Q, Jin X, Cui X. FFR_FD: Effective and fast detection of DeepFakes via feature point defects. Inf Sci. 2022;596:472–88.
27. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC. The deepfake detection challenge (dfdc) dataset. ArXiv. 2020. https://doi.org/10.48550/arXiv.2006.07397.
28. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. In: *Proceedings of the IEEE/CVF international conference on computer vision* 2019;1–11.
29. Jiang L, Li R, Wu W, Qian C, Loy CC. "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898. 2020.
30. Bhattacharyya C, Wang H, Zhang F, Kim S, Zhu X. "Diffusion Deepfake. arXiv. 2024. https://doi.org/10.48550/arXiv.2404.01579.
31. Li Y, Yang X, Sun P, Qi H, Lyu S. "Celeb-df: a large-scale challenging dataset for deepfake forensics." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020;3207–3216.