Detection of Attributes Manipulation in Deep fake Images using Hybrid Learning Technique to classify Indigenous and Forged Images

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

Signal Processing and Digital Design

Submitted by

Kanwardeep Singh Gahlot

(2K21/SPD/05)

Under the supervision of

Dr. Rajesh Rohila (Prof, ECE Dept.)

Mr. Rahul Thakur (Asst Prof, ECE Dept.)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi -110042

MAY 2023

M. Tech(Signal Processing and Digital Design) Kanwardeep Singh Gahlot 2023

CONTENTS

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I **Kanwardeep Singh Gahlot** student of M.Tech (Signal Processing and Digital Design), hereby declare that the project Dissertation titled "**Detection of Attributes Manipulation in Deep fake image using Hybrid Learning technique to classify indigenous and forged image"** which is submitted by me to the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **Kanwardeep Singh Gahlot**
Date: 31st May 2023

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Report titled "**Detection of Attributes Manipulation in Deep fake image using Hybrid Learning technique to classify indigenous and forged image**" which is submitted by **Kanwardeep Singh Gahlot**, **2K21/SPD/05** of Electronics and Communication Department, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 31st May 2023

Dr. Rajesh Rohilla
SUPERVISOR

Place: Delhi
Date: 31st May 2023

Mr. Rahul Thakur
SUPERVISOR

ACKNOWLEDGEMENT

A successful project can never be prepared by the efforts of the person to whom the project is assigned, but it also demands the help and guardianship of people who helped in completion of the project.
I would like to thank all those people who have helped me in this research and inspired me during my study.
With profound sense of gratitude, I thank Dr. Rajesh Rohilla and Mr. Rahul Thakur, my Research Guide, for their encouragement, support, patience and his guidance in this research work.
Furthermore, I would also like to thank the Head of the Department, Electronics and Communication, Prof O. P. Verma who gave me the permission to use all required equipment and the necessary format to complete the report.
I take immense delight in extending my acknowledgement to my family and friends who have helped me throughout this research work.

Kanwardeep Singh Gahlot

ABSTRACT

Deepfake technology employs artificial intelligence (AI) algorithms to create manipulated photographs and videos that are indistinguishable from authentic ones. Generative adversarial networks (GANs), a type of deep-learning algorithm, are driving the development of deepfake, which have the potential to compromise individual privacy, as they can be used to create pornographic content by superimposing images.As a result, digital media, including news broadcasts, online video clips, and live streams, are experiencing trust issues. Recently, synthetic image creation and manipulation methods have improved, enabling the creation of realistic fake face images. Despite the emergence of certain deep learning-based image forensic techniques, it is still challenging to differentiate between manipulated and genuine photos generated by modern techniques such as face2face, deep fake, and face swap. Current methods require substantial data input, high computational power, and are time-consuming. To overcome these challenges, we propose a novel Hybrid Learning model that ranks features to detect fake and genuine images using less computational power and time, while improving accuracy to 99 % and MCC score to 98.87 additionally reducing computing time by 40 seconds

Keywords: ***Deepfake, Artificial Intelligence, Generative Adversarial Networks, Hybrid Learning***

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full form
GAN	Generative Adversarial Network
KPCA	Kernel Principal Component Analysis
PCA	Principal Component Analysis
NNMF	Non Negative Matrix Factorization
AI	Artifical Intelligence
DIF	Digital Image Forensics
CNN	Convolutional Neural Network

CHAPTER 1
INTRODUCTION

1.1 Overview

In recent years, digital image alteration has become more widespread, making it challenging to determine the veracity and integrity of photographs. It can be done quickly to change a picture using techniques like cutting, cloning, and re sizing, which makes validation difficult. As a result, the study of digital image forensics (DIF) has grown to be a vital sector of knowledge devoted to confirming the legitimacy and dependability of digital information. DIF techniques are used by forensic professionals in their investigations to guarantee the reliability of digital evidence.

A recently developed technique called "Deepfake" uses artificial intelligence (AI) technology to manipulate facial videos. In order to create a new film that appears incredibly realistic and credible, it includes using photos and videos as input and fusing them together through a generative adversarial network.

1.1.1 What is Deepfake?

Deepfake is used to create fake films that can represent people talking and doing things they have never actually done by changing their faces in the footage. By influencing elections and discrediting politicians, Deepfake has the ability to incite conflicts between nations. Additionally, this method is currently being utilized to smear powerful politicians and celebrities by changing their appearances to those of other people. If exploited unethically, this could result in serious problems.

Even though the latest Generative Adversarial Networks (GAN) have greatly increased the development of Deepfake, some cutting-edge techniques may greatly enhance detection. [1]

There are majorly five types of Deepfake. These include:

1. **Textual Deepfake –** The development of textual Deepfake is made possible by the more advanced AI and technologies of the modern era. With the help of these developments, AI systems can now produce written material including articles, poems, blogs, and other types of writing. Software programme are capable of mimicking human thought, adapting to various situations, and using data to create literary works. [2]
2. **Deepfake Video –** The most common type of deepfake films are convincingly realistic videos produced with the use of artificial intelligence and video editing technology. Several software tools allow users to swap out a person's face for another person's in a video by making photographs available, thereby doing a face swap. Unfortunately, people often use these kinds of movies for malevolent purposes. [3]
3. **Deepfake Image -** Deepfake photos may be produced by nearly anyone and are frequently found online. People can easily view the various public figure images that are readily available online. Deepfake photos are produced in a manner similar to how applications and programme are used to create Deepfake films. Similar to the methods used in film editing, editing software is used to superimpose faces onto various bodies.. [4]
4. **Deepfake Audio -** AI is also capable of creating and altering audio content. A person's voice, including accent and tone, can be accurately reproduced. Algorithms are used by software programme to analyse already-existing voice samples and produce completely new ones that can say different things. When more recordings of the person being imitated are provided for reference, the accuracy of a Deepfake audio is improved. [5]
5. **Live Deepfake –** Deepfake technology has advanced quickly it has progressed to the point where it can currently create clones in a range of forms, including audio and video. Unfortunately, this advancement has made it easier for thieves and hackers to get through security precautions like voice-based verification. [6]

1.1.2 Generation of Deepfake

The development of Generative Adversarial Networks (GANs) has made it possible

to produce models that closely mimic the distribution of training data. The generator and discriminator are the two participants in an iterative min-max game that is involved in the process. The generator seeks to trick the discriminator by producing realistic examples, while the discriminator is trained iteratively to distinguish between genuine and created cases.

In recent times, GANs have also been used in adversarial assaults, particularly in CF (Copy-Move Forgery) attacks. For instance, Kim et al. developed a median filtering CF attack that effectively eliminates the evidence of median filter images by using GAN networks. Benford's law was also used by Bonettini et al. to discern between edited and produced photographs. Benford's law investigates how the leading DCT coefficients, offering a way to spot photos that have been altered. [7]

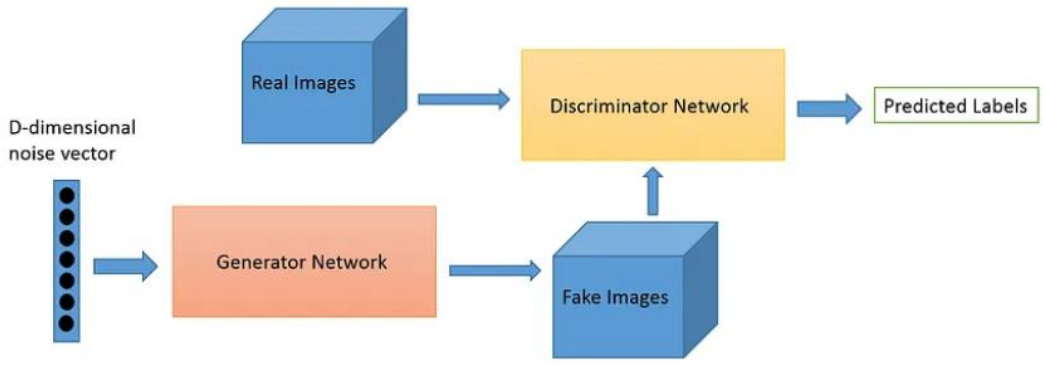

Fig 1.1: Flow diagram of Generative Adversarial Network [8]

1.1.3 Deepfake and it’s applications

Deepfake, a combination of "deep learning" and "fake," refers to extremely realistic videos that have been digitally altered to depict people saying and acting in ways that have never actually happened. These films use neural networks to learn and imitate a person's facial expressions, gestures, speech, and intonations by analysing large datasets. To swap the faces of two people, a deep learning system is trained by being fed footage of the people. Deepfakes basically swap out one person's face in a video for another person's by using facial mapping technology and artificial intelligence [9].

Movies, education, digital communications, gaming, entertainment, social media, healthcare, material science, and numerous business sectors including fashion and e-commerce are just a few of the domains where deepfake technology has found useful uses.

The Deepfake technology has various advantages in the film industry. For actors who are unable to talk due to illness, it can be utilised to create synthetic voices. It can also be used to update film footage rather than reshoot scenes. With the aid of this technology, filmmakers can reproduce famous sequences, revive performers who have passed away, add cutting-edge special effects and facial editing in post-production, and raise the calibre of amateur videos to a professional level. Deepfake technology also enables automatic and lifelike voice dubbing in any language, improving the enjoyment of films and educational content by various audiences. As an illustration, consider the global malaria awareness campaign that included David Beckham in 2019 and used speech and visual alteration technologies to make him appear bilingual and break down language boundaries.

,

1.1.4 Major factors that pose as risk in use of Deepfake

Deepfake technology introduces various risks and difficulties

1. **Disinformation and False News –** Deepfake have the ability to create highly realistic yet fabricated content such as videos, images, and audio recordings. This presents a significant risk of spreading disinformation and false news. Deepfake can manipulate public opinion, fabricate narratives, and defame individuals
2. **Social Engineering and Fraud -** Deepfake can be employed for social engineering attacks and fraudulent activities. They can be used to impersonate individuals, including influential figures, politicians, or business leaders, with the intention to deceive and manipulate others for personal gain or malicious purposes.
3. **Privacy Concerns: –** Deepfake technology enables the creation of non-consensual explicit content by superimposing someone's face onto explicit images or videos without their consent. This raises serious privacy concerns

and can result in harassment, blackmail, or damage to one's reputation.

4. **Trust and Authenticity: -** Deepfakes undermine trust and authenticity in digital media. As deepfake techniques become more advanced and harder to detect, distinguishing between genuine and manipulated content becomes increasingly challenging. This erosion of trust can have far-reaching consequences in domains such as journalism, law enforcement, and personal relationships.
5. **Impact on Elections and Politics -** Deepfake possess the potential to disrupt democratic processes by spreading false information or manipulating political events. Deepfake can be used to misrepresent statements or actions of political candidates, leading to public confusion and a loss of trust.
6. **Cybersecurity Threats: -** Deepfakes can be utilized as tools in cybersecurity attacks. To illustrate, individuals with malicious intent have the ability to generate manipulated audio or video, known as deepfakes, in order to mislead people or obtain unauthorized entry into systems or access sensitive data..

Addressing these risks requires a comprehensive approach, including advancements in deepfake detection and authentication technologies, public awareness and education about deepfake, the establishment of policy and legal frameworks, and the responsible use of AI technologies. [11]

1.2 Literature review

In recent years, digital picture alteration has become quite widespread. So it is difficult to verify the integrity and authenticity of photographs because it is so easy to edit, clone, and resize an image. In their forensic investigations, forensic specialists can also use digital photographs. Digital image forensics (DIF), which focuses on confirming the integrity and authenticity of digital files, has grown to be a crucial field of knowledge in this context. Generative adversarial networks (GANs) can learn deep representations without relying heavily on labeled training data. This is accomplished through a competitive procedure that involves two networks producing backpropagation signals. GANs possess a broad spectrum of applications, such as generating images, modifying images based on meaning, transferring artistic styles, enhancing image resolution, and performing classification tasks. These applications are made possible by GANs' ability to acquire meaningful representations through

learning.. The objective is to provide the signal processing community with a comprehensive understanding of GANs, employing familiar analogies and concepts whenever possible. We discuss different approaches for constructing and training GANs while addressing remaining challenges in their theoretical understanding and practical implementation. Convolutional Neural Network (CNN) features are used by (Guera et al. 2018) to automatically detect deepfake films, and then a Recurrent Neural Network (RNN) is used to distinguish between authentic and fake images. A 96.9% validation accuracy was achieved using 300 deepfake videos in the experiment. (Ding et al.2021) introduced a novel anti-forensic GAN algorithm with additional supervision modules for improving picture visual quality.

The suggested technique effectively creates anti-forensics videos with pleasing visual quality and no discernible artifacts. For the datasets Face Forensics ++, Celeb-DF, and DFDC, Architecture reported accuracy of 98.95%, . 97.17%, and 97.24 using XceptionNet, respectively. As a result, XceptionNet is preferred over DenseNet and ResNet because of its superior accuracy for Face Forensics++, which was 96.72% and 96.55%, respectively .

Table 1: Types of data sets available

S. No	Data set	Year	Information	Remarks
1	NIST Nimble	2016-17	Measurements of the detection and localization of automated imagery (picture and video) modification. Supports GAN	AlexNET Model Proposed **98.178%** accuracy[12]
2	CASIA V2	2016	It is a dataset for forgery classification. Dataset consists of 4,795 images, with 1,701 being genuine and 3,274 being forged.	CNN + DWT proposed **97.2** % accuracy [13]
3	DeeperForensics-1.0:	2020	Larger Dataset than already present	The accuracy reported was **64.1 %** [14]

4	FFIW10K Dataset	2021	The computer vision community advances through the use of challenging datasets. In order to give a better benchmark and aid in identifying the circumstances in which present algorithms fail, FFIW10K was developed in the intention of encouraging additional research efforts.	**69.4%** Accuracy Reported [15]

1.3 Research Gap

In the methods listed under the detection of Deepfake, there are two issues: requirement of high computation power and low classification accuracy. These two drawbacks have the following consequences:

1. When one trains deep learning models based on Deepfake data, the models required high computation power as there are features in range of 512-1024, and most of the features among them are redundant and does not significantly contribute to classification, instead they increase the computation power and computation time which may effect significantly in real time situation. As a result, the important features test instances are wrongly interpreted more often as compared to the majority ones. Therefore, while handling binary classification cases (like in the case of Deepfake data), the model is highly sensitive (if the image is of the majority class) and depicts low specificity (when indigenous image are in minority). This is a clear indication of biasedness for the majority class.

Remedy for Computational power : The general method used to remove the issue of high computational power can be done in two ways:

a) Features ranking – The features of image dataset of every class in the training data is to be done using an deep neural network and optimizing the features based on ranking with Chi square2 scores and consideration of only important features based on threshold value of score .

b) Features classification – The features classification to be done using a machine learning algorithm only with selected features from a range of features set of 512 -

1024 and testing the number of features to enhance accuracy and training the models on a selective balance of features of each class in the training data.

Remedy for low accuracy rate: An ensemble model is used for prediction or classification. The most significant advantage of using ensembles is to enhance the average prediction performance over any contributing member in the ensemble.

1.4 Research Objective

This dissertation has the following objectives:

1. To explore the impact of model hybrid learning and
    i. Ranking features for Deepfake detection can be accomplished by employing Kernel Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF).
    ii. To enhance the accuracy of classifying Deepfake image datasets, a hybrid learning model is proposed as the prediction model. The goal is to improve the classification accuracy in this context.
2. The aim is to enhance the classification accuracy specifically within this domain., specificity, F-score, Mathew Correlation Coefficient (MCC) and ROC curve.

1.5 Structural organization of the dissertation

The remaining part of the dissertation is arranged in the following manner: Chapter 2, describes the problem of Deepfake in detail. It also gives the solutions developed over the years through research to solve the forged images in Deepfake datasets . Further, it highlights the importance pre-processing of images and features ranking for Kernel PCA analysis. It also covers the aspects like dataset used and step by step execution of the proposed method. Chapter 2 describes the effects proposed method to enhance segmentation. Chapter 3 encompasses the evaluation and validation of the method's performance parameters. Additionally, it highlights the advantages and limitations of the work. On the other hand, Chapter 4 serves as the conclusion, addressing future prospects and aspects of the research.

CHAPTER 2

PROPOSED HYBRID LEARNING TECHNIQUE FOR DETECTION OF DEEPFAKE USING KERNEL PCA AND NON NEGATIVE MATRIX FACTORIZATION

2.1 Introduction

This chapter is focused on three major factors:

1. Need for augmentation of Deepfake image data set.
2. The importance of Kernel- PCA for deep fake detection.
3. Introduction and step by step execution of the proposed methodology.

Deepfake is a serious issue since it uses techniques like face swapping, attribute manipulation, face morphing, and pose transfer to alter actual photos. The datasets for Deepfake are unbalanced since more fake images are produced by attribute manipulation for each real image. As a result, a single real image is used to create hundreds of phoney ones. The prediction model trained on such uneven datasets may produce biased predictions if this imbalance is not corrected during pre-processing. This bias describes the model's propensity to favour samples from the majority of classes. In comparison to the minority class, the model learns more about the majority class because there are more examples from this group. When learning crucial information about the minority class is the goal, this bias presents a considerable barrier. Resampling techniques can be used to balance the dataset to address this problem. Additionally, the accuracy of the prediction, which could have been hampered by the dataset imbalance, can be significantly improved by utilizing a classifier ensemble. In order to solve this issue, an augmentation pre-processing technique is used to produce more images for the real image collection.

The significance of Kernel Principal Component Analysis is emphasized in the latter portion of the chapter. Kernel PCA, an extension of the traditional PCA method that facilitates non-linear analysis, is introduced. PCA is widely employed to convert

high-dimensional data into a lower-dimensional representation while retaining critical information.

Kernel PCA starts by transforming the input data into a feature space with higher dimensions using a non-linear function called a kernel. This kernel function quantifies the similarity between data points in the original input space. Through this mapping process, Kernel PCA can effectively capture non-linear connections among variables that might go unnoticed by linear PCA.

After the data is transformed into the higher-dimensional space, the conventional PCA method is employed to calculate the principal components. These principal components indicate the directions of the highest variance within the transformed feature space. Nevertheless, due to the utilization of a kernel function for the transformation, the principal components are expressed in relation to the original input space.

Kernel PCA is particularly useful when dealing with complex datasets that exhibit non-linear structures or when the linear PCA fails to provide meaningful representations. It has applications in various domains, including image processing, bio informatics, and natural language processing, where non-linear relationships are often present.

2.2 Impact of imbalance on a prediction model

This section illustrates the details of what exactly is class imbalance issue, its remedies and why the Kernel Level PCA and Non Negative Matrix factorization techniques are used for the research.

2.2.1 Class Imbalance

As a result of the datasets' lopsided distribution of data samples, unbalanced classification is a problem. It has the following characteristics:

1. **Class overlapping:** Data instances from different classes cross paths (Fig. 2.1 a). The classifiers struggle to distinguish between several classes accurately in such cases. As a result, samples related to the minority class are mistakenly classified as belonging to the majority class.

2. **Small sample size:**In practise, gathering enough information for imbalanced datasets is rather challenging. The imbalance ratios in the datasets can be balanced to

reduce misclassification error as a solution to this problem.

3. **Small disjuncts:** As seen in Fig. 2.1 b, minority class data instances are dispersed throughout several feature spaces. This makes the classification stage more difficult. The size of samples from two distinct classes can be distinguished in a meaningful way (big ratio of imbalance). The classifiers may interpret a small number of minority class data examples as aberrations, leading to a significant misclassification rate for the minority class. The impact of the class imbalance problem grows as the volume of the data increases. When working with real-world datasets, class imbalance arises when one class, referred to as the minority class, contains fewer instances compared to the other class, known as the majority class, which has a greater number of instances.. Without pre-processing, building an optimum model using traditional data mining and machine learning techniques is an impossible undertaking. Pre-processing primary purpose is to balance the datasets.

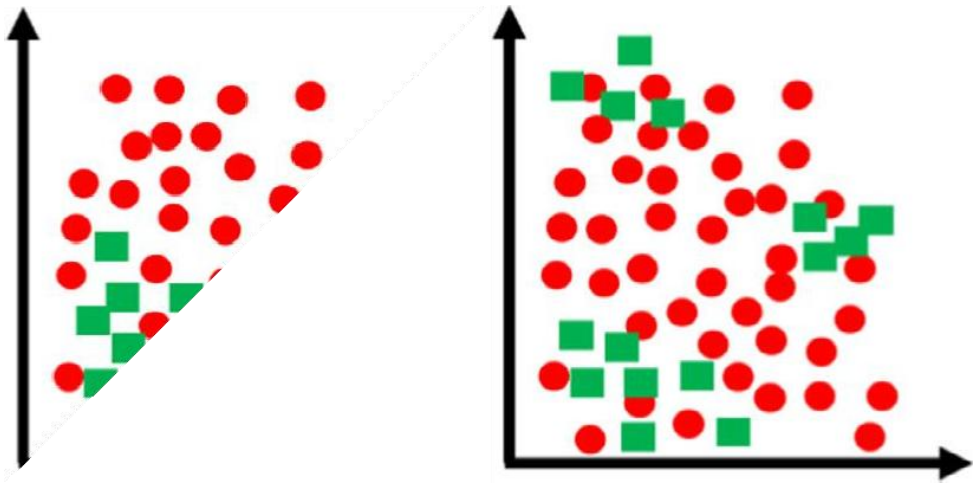

Fig. 2.1 a) Example of class overlapping b) Example of small disjuncts

If the issue of class imbalance is ignored, the built-in models or learning algorithms may end up being overrun by the majority class and hence disregard the minority class. Let's take a look at a two-class dataset with a 99% imbalance ratio, where the dominant class makes up 99% of the data set and the minority class only makes up 1%.

The learning approach groups the entire sample set into the majority class, which has an error rate of 1%, in order to minimize the error rate. In such a case, the minority class cases are crucial and must be acknowledged as being incorrectly classified.

The methods to solve this imbalance issue are classified as follows:

1. Algorithmic-level techniques
2. Data-level techniques
3. Cost-sensitive techniques
4. Classifier ensembles

Prior to developing the classifier, the data-level strategies focus on preprocessing the imbalanced datasets. In this method, the preparation of the data and the training of the classifier can be done individually and without correlation. Resampling techniques serve as the foundation for data pre-processing approaches. Before the model moves on to the training phase, it is completed.

2.2.2 Solutions for class imbalance

There are two ways to remove the class imbalance issue on the basis of data, algorithms, and cost sensitivity

a. Augmented Data: Deep networks often require a substantial amount of training data to achieve good performance. Image augmentation is commonly employed to enhance the effectiveness of deep networks and enable the creation of robust image classifiers with limited training data. By employing diverse processing techniques like random rotation, shifts, shear, flips, and others, image augmentation artificially generates additional training images. Many academic disciplines use deep convolutional networks to conduct Computer 17 Vision tasks in an effort to outperform current benchmarks. One of the hardest problems is enhancing these models' capacity for generalisation. When a model is tested on data that has already been seen (training data) vs data that has never been seen before (testing data), generalizability is used to describe the performance difference. Poorly generalizable models have overfit the training set of data. To detect overfitting, a common approach is to plot the training and validation accuracy at each epoch during the training process. Generalizability, the ability of models to perform well on unseen testing data compared to the training data they have already encountered, is a challenging aspect to improve. Models that lack generalizability have excessively fit the training set. By plotting the training and validation accuracy at each epoch, overfitting can be identified. Models that exhibit poor generalizability are prone to

overfitting the training set.

Types of image augmentation:

A. Translation - The act of moving an image's position horizontally or vertically within the image plane is referred to as translation. It entails shifting the pixels of the image in a particular direction without changing their look or information.

B. Rotation - An image is rotated when its orientation is altered by a certain amount. It can be applied to achieve a specific viewing angle or align tilted photos.

C. Shearing - Shearing bends the image along one axis while holding the other axis fixed, causing distortion. It can be applied to create unusual effects or to fix perspective distortion.

D. Scaling - Scaling either increases or decreases the size of an image. This transformation can be applied to photos to zoom in or out or resize them to a particular dimension.

E. Affine Transformation - Translation, rotation, scaling, and shearing are all combined in an affine transformation. They have the ability to bend an image irregularly while maintaining parallel lines and the relative sizes of items.

F. Perspective Transformation - Transforming the perspective of an image: Transforming the perspective of an image reproduces the impression of viewing an image from a different vantage point or camera angle. They can be used to create 3D-like effects or to fix perspective projection distortion.

Fig 2.2 Original Images

Fig 2.3 Augmented Images

Histogram Equalization : We employ a technique that can enhance image details to emphasize differences between actual and false photos, which was inspired by the difference between real and fake facial saliency maps. The two primary categories of conventional picture enhancement methods are the frequency domain and spatial domain. The image is immediately processed by spatial domain improvement methods, such as grayscale transformation and histogram equalization. The frequency-domain enhancement algorithm mostly works with images that fall under a specific range of image variation, such as the Fourier transform and Wavelet transform. In recent years, guided filter, an edge-preserving filtering method, has seen widespread applications. [16] The frequency-domain enhancement algorithm mostly works with images that fall under a specific range of image variation, such as Fourier transform and Wavelet transform [17]. In recent years, guided filter, an edge-preserving filtering method, has seen widespread applications. In this work, we are implementing Exposure based Sub Image Histogram Equalization for colored images to enhance the image.[18] Images that lack contrast fail to utilize the full dynamic range available. When the histogram bins of photographs are concentrated towards the brighter portion, they exhibit high-intensity exposure, whereas bins concentrated towards the darker levels indicate low-intensity exposure. Images can be roughly classified as underexposed or overexposed based on the strength of their exposure. In this section, the ESIHE algorithm is introduced, which consists of three steps: determining the exposure threshold, histogram clipping, and histogram subdividing a and equalizing.

$$exposure = \frac{1}{L}\frac{\sum_{k=1}^{L} h(k)k}{\sum_{k=1}^{L} h(k)} \quad \text{……..eq 2.1}$$

The histogram of an image, denoted as h(k), represents the frequency distribution of gray levels. The total number of gray levels in the image is represented by L.

$$X_a = L(1 - exposure) \quad \text{……..eq 2.2}$$

For an image with a dynamic range of 0 to L, this parameter takes a value greater than or less than L/2 (gray level) when the exposure value is greater than or less than 0.5, respectively.

$$T_c = \frac{1}{L}\sum_{k=1}^{L} h(k)$$

$$h_c(k) = T_c \quad \text{for } h(k) \geqslant T_c$$ .....eq 2.3

In this method of histogram clipping, the original histogram is denoted as h(k), while the clipped histogram is represented as hc(k). This approach offers computational efficiency and consumes less time compared to alternative methods.

Figure 2.4 represents ESIHE enhanced image and original image, ESIHE could be used a first stage segmentation if there is any change observed in particular area with enormous changes in contrast and brightness.

Fig 2.4 Original vs ESIHE enhanced image

2.3 Adopted methodology

The steps for the method used for the prediction of deep fake using Kernel PCA is shown below:

i) Selecting the datasets.

ii) Pre-processing the datasets to balance classes to improve quality of dataset for indigenous images.

iii) Extracting the features using ResNet 50 deep neural network.

iv) Analysis of features and reducing dimension using Kernel Principal Component and Non Negative Matrix factorization.

v) Ranking of features based on Chi square 2 scores and selecting features on basis of threshold score.

vi) Classification using machine learning based on selected features.

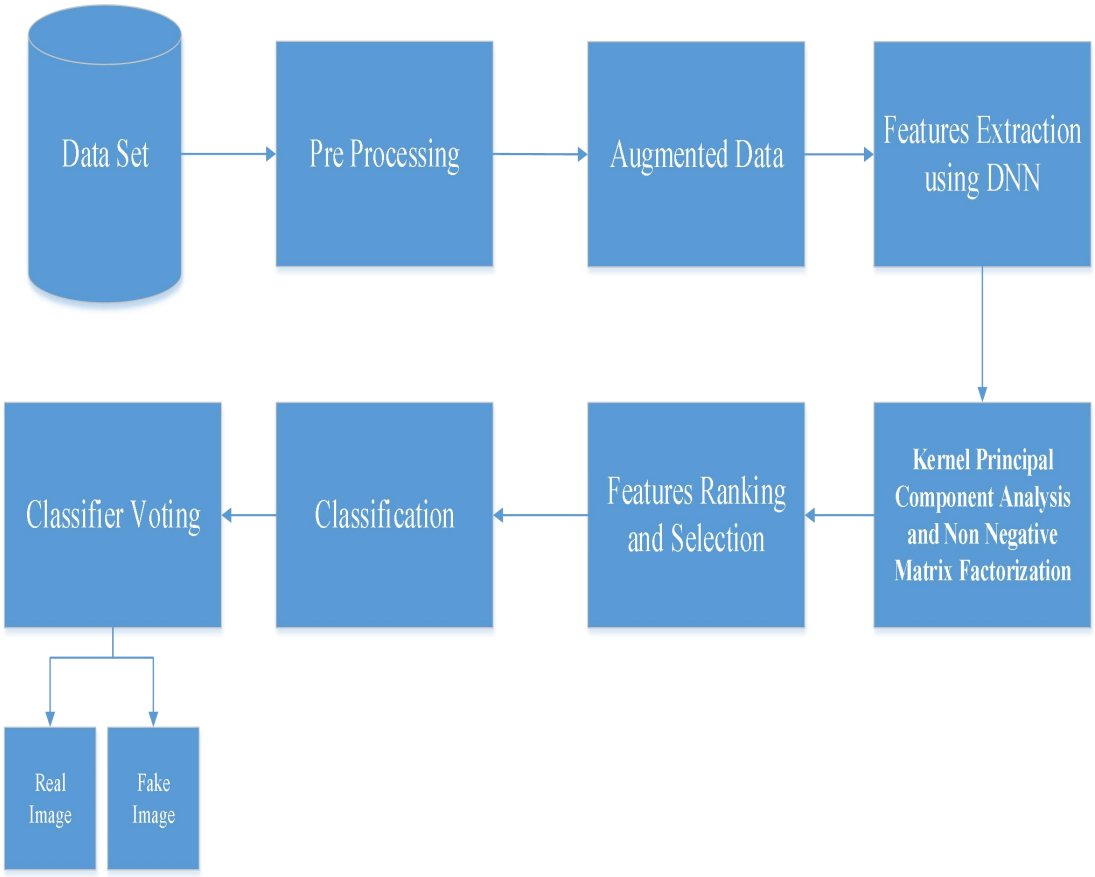

Fig. 2.5 Block diagram of generalized methodology for detection of Deepfake.

2.4 Reason of choosing Hybrid learning technique over other methodology

This section Deep learning and Machine learning in detail. It also gives the reason behind the preference of these two techniques over other detection technique. Hybrid learning using a combination of these two is also discussed briefly.

2.4.1 Deep learning network for features extraction?

In this work, we extract image characteristics utilising state-of-the-art architectures for image classification such Xception net, Resnet 50, and Google net. According to Table 2, which is based on the network, each deep neural network's number of connections, output layers in the image, and features are discussed in table 2.

Table 2: Details of Deep Neural Network for feature extraction

Features	Xception Net	ResNet50	GoogLe Net
No. Of Layer	170	177	144
No. Of Connection	181	192	170
Image input	299*299*3	224*224*3	224*224*3
Output Layer	Global Averaging Pool	Global Averaging Pool	Global Averaging Pool
No of Features	1024	2048	512

Design of a convolutional neural network Microsoft Research originally introduced ResNet-50, formerly known as Residual Network-50, in 2015. It belongs to the ResNet family, which has gained a lot of acceptance and recognition among computer vision experts.

The 50-layer network comprises convolutional layers, pooling layers, fully connected layers, and skip connections, also referred to as residual connections. The incorporation of skip connections enables the network to learn residual functions, addressing the degradation problem encountered when training deep networks. ResNet-50 consistently achieves remarkable outcomes when evaluated on challenging test datasets like ImageNet, which require advanced visual recognition capabilities. ResNet-50 is widely recognized for its exceptional performance in image classification tasks. Currently, it serves as a fundamental architecture in

various computer vision applications, including object identification, image segmentation, and image recognition tasks.Its architecture includes residual blocks, which serve as the network's basic building components. Convolutional layers that have been swiftly piled and connected together make up these remaining blocks. These connections are designed to make identity mappings simpler to comprehend, which will make it simpler to train and improve deeper models. ResNet-50 is a widely adopted and powerful neural network architecture in the field of computer vision due to its ability to achieve a remarkable balance between accuracy and computational complexity.. [16]

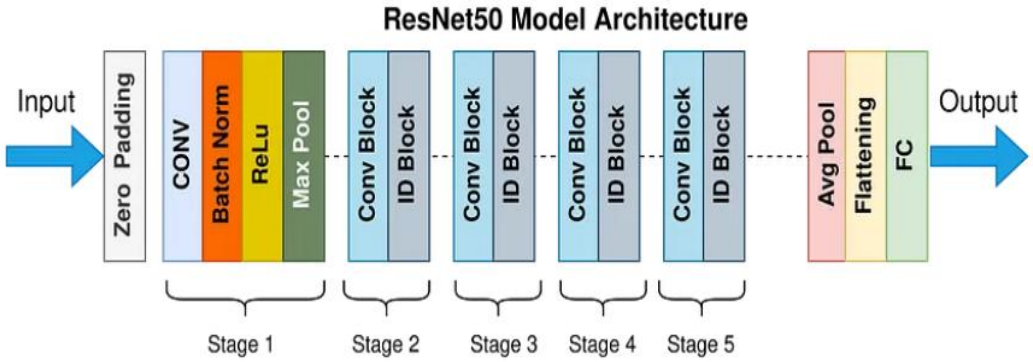

Fig. 2.6 Architecture of ResNET 50 Model Deepfake [17]

2.4.2 Why Machine learning for classification ?

Machine learning is employed in image classification because it can autonomously learn and extract significant patterns and features from vast quantities of image data. Unlike traditional approaches that relied on manual rule-based methods and handcrafted features, machine learning techniques offer the advantage of capturing complex patterns in a more efficient and effective manner. This eliminates the need for labor-intensive and time-consuming manual feature engineering. Machine learning algorithms, specifically convolutional neural networks (CNNs), have transformed the field of image classification by enabling the automatic learning of hierarchical representations directly from raw pixel data. The following are notable reasons for the widespread utilization of machine learning, particularly CNNs, in image classification:

2.4.2.1. Feature Extraction: Convolutional neural networks (CNNs) possess the capability to autonomously learn and extract pertinent features from images at various levels of abstraction. They are proficient in capturing both low-level features, such as edges and textures, and high-level features, such as shapes and objects. The hierarchical architecture of CNNs is purposefully designed to acquire complex patterns within the data in a step-by-step manner.

2.4.2.2. End-to-End Learning: Machine learning facilitates end-to-end learning, wherein the model simultaneously learns the feature representation and classification decision. This eliminates the requirement for manual feature engineering and empowers the model to automatically learn the most distinguishing features essential for the classification task.

2.4.2.3. Adaptability: Machine learning models can adapt to a wide variety of image classification tasks by training on labeled datasets. They can generalize from the training data to accurately classify unseen images, provided the training data is representative and diverse.

2.4.2.4. Scalability: Machine learning techniques can handle large-scale image classification tasks efficiently. With the availability of powerful hardware, such as GPUs, training deep learning models on massive datasets has become feasible, enabling robust and accurate image classification on a large scale.

2.4.2.5. Performance: Deep learning models, such as CNNs, have exhibited remarkable advancements in the realm of image classification tasks. They have surpassed traditional methodologies and achieved state-of-the-art accuracy on widely recognized benchmark datasets. Consequently, deep learning models have gained widespread acceptance and are extensively utilized in diverse applications.Machine learning for image classification has found applications in numerous fields, including medical imaging, autonomous driving, surveillance systems, facial recognition, and more. It allows for automated and efficient analysis of visual data, providing valuable insights and enabling advanced decision-making based on images.

2.4.3 Hybridization: Hybrid Learning

Hybrid learning, which combines both traditional machine learning methods and deep learning techniques, can be highly useful in various scenarios. The following

five major point, discuss why hybrid machine learning is advantageous:

2.4.3.1. Leveraging the strengths of both approaches: Traditional machine learning techniques are effective in handling structured data and performing feature engineering, while deep learning excels in processing unstructured data and extracting intricate patterns. By combining the two, hybrid models can leverage the strengths of each approach, leading to more accurate predictions and improved performance.

2.4.3.2. Enhanced feature representation: Deep learning models, including CNNs and RNNs, possess the capability to autonomously learn intricate features directly from raw data. These acquired features can subsequently be utilized as input for traditional machine learning algorithms, like SVMs or decision trees, resulting in an enhanced representation that enhances the final predictions.

2.4.3.3. Dealing with extensive and high-dimensional data: Deep learning models have a reputation for effectively managing large-scale and high-dimensional data, including images, videos, and natural language. By integrating deep learning techniques into hybrid models, it becomes possible to efficiently process and extract valuable information from such data, leading to enhanced accuracy and insightful predictions.

2.4.3.4. Transfer learning and pretraining: Deep learning models are often pretrained on large datasets, such as ImageNet, to learn general-purpose representations. These pretrained models can then be fine-tuned using specific datasets and combined with traditional machine learning techniques. This approach, known as transfer learning, can significantly reduce the need for large labeled datasets and speed up the training process, making it particularly useful in domains with limited data availability.

2.4.3.5. Interpretable and explainable results: Traditional machine learning models are often more interpretable and provide explicit insights into the decision-making process. By combining deep learning with traditional techniques, hybrid models can retain some interpretability while benefiting from the enhanced predictive power of deep learning. This is crucial in domains where model interpretability is essential, such as healthcare and finance.

Overall, hybrid machine learning models offer the advantages of both traditional and deep learning approaches, enabling more accurate predictions, improved feature representation, and enhanced scalability, while also providing interpretability when

required.

2.5 Proposed Methodology

This section gives the step by step process explanation of the method used

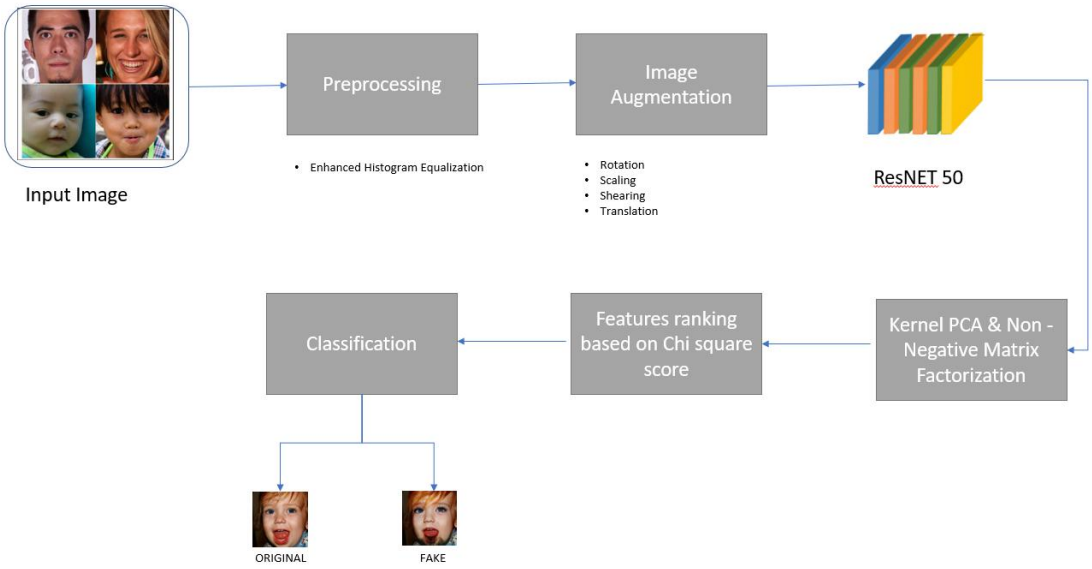

Fig. 2.7 Architecture of proposed methodology

2.5.1 Data acquisition

The Deep Fake Face Data set (DFFD) is an inclusive compilation of data that plays a crucial role in the detection and localization of manipulated faces. It encompasses four primary categories of facial manipulations: identity swapping, expression swapping, attribute manipulation, and fully synthesized faces. To construct this data set, cutting-edge techniques were employed to generate fabricated images representing each of these categories. The collected images and video frames consist of approximately 47.7% male subjects, 52.3% female subjects, and the majority of samples fall within the age range of 2-50 years. Both low-quality and high-quality images are included, encompassing variations in face size for both real and fake samples. Video clips from FaceForensics++ were utilized for facial identity and expression swapping. Two methods were employed to generate attribute-manipulated images: FaceAPP and StarGAN. FaceAPP, a consumer smartphone app, provides 28 filters that can modify specific facial attributes such as gender, age, hair, beard, and glasses. For each face in the FFHQ data set, three corresponding fake images were

generated: two with a single randomly selected manipulation filter and one with multiple manipulation filters.

Figure 2.8: Fake images generated using attribute manipulation, expression swap, and identity swap [18]

Figure 2.9: Original Image data set [18]

2.5.2 Kernel Level Principal Component Analysis

Principal Component is a method for removing noise from a signal and decomposing it into various Principal components, each with a varying level of significance depending on the embedding dimension. PCA is a model-free and exploratory technique that may be used in real-time or batch mode if the class of series is not too vast and well-defined so that the rule for selecting appropriate parameters can be fixed. PCA is a technique employed to decrease the dimensionality of extensive datasets. It accomplishes this by transforming a larger set of variables into a smaller one while retaining the majority of the pertinent information. The aim of dimensionality reduction is to simplify data by reducing the number of variables. This process entails sacrificing a certain degree of accuracy for the sake of simplicity. The advantage lies in the fact that smaller datasets are more manageable for exploration, visualization, and analysis. By eliminating extraneous variables, the processing of data points becomes more efficient and rapid for machine learning algorithms.

Kernel Principal Component Analysis (KPCA) is a technique used in machine learning to reduce the dimensionality of nonlinear datasets. It extends the classical Principal Component Analysis (PCA) algorithm, which is limited to linearly separable data. KPCA employs a kernel function to map the data into a higher-dimensional feature space, where linear methods like PCA can effectively capture nonlinear relationships. By using the kernel function, KPCA transforms the input data, allowing for the identification of significant components in a nonlinear context. This enables the analysis of complex and nonlinear patterns within the data, making it useful for tasks like data visualization, clustering, and classification.

KPCA offers advantages over traditional PCA by accommodating nonlinear relationships between features. This capability is particularly valuable in domains like image or speech recognition. Additionally, KPCA can handle high-dimensional datasets with numerous features while preserving critical information and reducing dimensionality. However, KPCA does have limitations. Selecting an appropriate kernel function and its associated parameters can be challenging and time-consuming. Furthermore, for large datasets, KPCA can be computationally expensive since it requires calculating the kernel matrix for all pairs of data points.

$$\pi(x) = \sum_{j=1}^{m} v_j y_j = \sum_{j=1}^{m} y_j \sum_{i=1}^{m} a_{ji} \emptyset (x_i) \text{ .....eq (1)}$$

$$v_j = \sum_{i=1}^{m} a_{ji} \emptyset (x_i) \text{ .....eq (2)}$$

y_j is the coordinate of φ(x) along one of the feature space axes v_j

Algorithm to calculate KPCA:

Algorithm 1: Kernel Level Principal Component Analysis

1: To start we select a kernel function, denoted as k(x_i, x_j),and then proceed to select any transformation T, that maps the data to higher-dimensional space.

2:To centre our kernel matrix, which is equivalent to compute this matrix. Consequently, we will generate a kernel matrix which is formed by applying the kernel function to every pair of data points.

$$K\_new = K - 2(I)K + (I)K(I)$$

3. Next, we proceed to compute the eigen vectors and eigen values of the centered kernel matrix.

4. We will specify the desired number of dimensions for our reduce datasets, denoted as "m". Then, we will select the first "m" eigen vectors and combine them into a single matrix

.

2.5.3 Non Negative Matrix Factorization

Non-negative matrix factorization (NMF) is a dimensionality reduction technique used for data analysis and feature extraction. It is particularly suitable for datasets where all the values are non-negative, such as non-negative images or text data. The goal of NMF is to represent a given non-negative matrix as the product of two non-negative matrices of lower rank. By doing so, it aims to uncover the underlying latent factors or patterns in the original data. These latent factors are represented by the columns of the factor matrices.

Mathematically, given an input non-negative matrix X, NMF seeks to find matrices

W and H, such that X ≈ WH, where W represents the basis matrix and H represents the coefficient matrix. The columns of W and H capture different aspects of the data and can be interpreted as representative patterns or features. The NMF algorithm iteratively updates the factor matrices W and H by minimizing the reconstruction error between X and the approximation WH. This is typically achieved by employing optimization techniques like gradient descent or multiplicative updates.

NMF has found applications in various domains, including image processing, document clustering, topic modeling, and recommended systems. It offers advantages such as interoperability of the resulting factors, sparsity in the factorization, and the ability to handle non-negative data effectively.

It is important to note that NMF is a form of unsupervised learning, meaning it does not require labeled data for training. Instead, it focuses on extracting meaningful patterns solely from the input data itself.

$$A_{m\,x\,n} = W_{m\,x\,k}\ H_{k\,x\,n} \quad \text{.....eq (3)}$$

Where,

A - original input matrix

W - feature matrix

H - coefficient matrix

k - low rank approximation

Algorithm to calculate NNMF:

Algorithm 2: Algorithm to Non Negative Matrix Factorization

1: Initialize:
a.) Choose the desired number or components or latent factors, denoted as "k".
b.) Initialize the factor matrices W and H with random non-negative values or other initialization methods

2:Update W and H:
a.) Repeat until convergence or a specified number of iterations:
b.) Update the matrix W by minimizing the reconstruction error between X and WH, while enforcing non-negativity constraints.
c.)Update the matrix H by minimizing the same reconstruction error, subject to non-negativity constraints.

3. Convergence:

Check for convergence by monitoring the reconstruction error or other convergence criteria.

4. Obtain the NMF representation:
Once the algorithm has converged, the final factor matrices W and H represent the NMF representation of the original data X.

2.5.4 Features Ranking

Several features were extracted, ranging from time domain and spectral domain such as power bandwidth, approximate entropy, peak 2 peak, complexity, mobility, etc. were extracted and the features were tested based Chi Square test.

Chi Square Test: The Chi-square test is used to examine if two variables are independent. It can be used to determine whether there is a correlation between two discrete variables . The bigger the variance between the two variables, the less the correlation between them, and the stronger the independence, the higher the Chi-square value. Extracted fourteen features were ranked based on the scores of the Chi-square test Since the problem is of a binary classification it examined differences between categories from a random sample to judge the goodness of fit. Table 2 depicts the Chi-Square test obtained by all the features for classification between binary states i.e., Mental Arithmetic and Rest state statistical features such as feature 2, feature 3, feature 5, feature 10 and feature 11 were selected for training. Statistical features such as different order moments, variance, and range of the signal showed dominance over other features such as power bandwidth, Lyapunov exponent and kurtosis. The formula for calculating the chi-square test statistic depends on the type of data and the specific hypothesis being tested.

$$\chi^2 = \Sigma [(O - E)^2 / E] \text{ ……..eq (4)}$$

Where:

χ^2 is the chi-square test statistic

Σ represents the summation symbol

O represents the observed frequencies in each cell of the contingency table

E represents the expected frequencies in each cell of the contingency table

The chi-square test statistic adheres to a chi-square distribution with (r-1)(c-1) degrees of freedom, where r represents the number of rows in the contingency table and c denotes the number of columns. Once the chi-square test statistic is computed, it can be compared against the critical value obtained from the chi-square distribution or utilized to calculate the p-value. These comparisons help ascertain the statistical significance of the association between the variables.

CHAPTER 3
RESULTS AND DISCUSSION

3.1 Introduction

In this chapter, the performance parameters employed to assess the accuracy of the methodology are discussed in a comprehensive manner. After completing the steps of data balancing and prediction, the validity and efficiency of the process are evaluated. The performance of the model employed in this study relies on five key parameters: F-measure/F-score, accuracy, specificity, sensitivity and Matthew Correlation Coefficient (MCC). Additionally, the effectiveness of the proposed model is tested using performance parameter analysis.

3.2 Performance parameters

The performance of the proposed model is evaluated using five evaluation metrics, namely:

a) Matthew Correlation Coefficient (MCC)
b) F-measure/F-score
c) Accuracy
d) Specificity
e) Sensitivty

The conventional accuracy metric is inadequate for capturing the true performance of a model when dealing with imbalanced data. Hence, the balanced accuracy metric is employed to provide a more accurate representation of model behavior. The following metrics are calculated to assess performance: accuracy, balanced accuracy, sensitivity, and specificity :

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{.....eq (5)}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TP} + \text{FN}} \quad \text{.....eq (6)}$$

$$Accuracy_{bal} = \frac{Sensitivity + Specificity}{2} \quad \text{.....eq (7)}$$

$$\text{F} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \text{.....eq (8)}$$

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \text{.....eq (9)}$$

where TP - True Positive

FP - False Positive

TN - True Negative

FN - False Negative

F - F-score/F-measure. It returns a value between 0 and 1 where 0 means worst prediction and 1 indicates perfect prediction.

MCC – It is usually between -1 and 1 where -1 means worst prediction and 1 indicates perfect prediction.

In order to validate its effectiveness, the proposed model is compared with five other earlier models. The first two models and work with Face forensics++ data using CNN and LRE technique 94.33% and 98.90 % accuracy respectively. The GANs celeba were with CNN to obtain 99.91% accuracy but limited to generated deepfake. The CFFN used to remove model biasedness yielded significant improved accuracy of 98.04%. Table 3.1 shown below gives the detailed description of the effectiveness of the proposed hybrid resampling compared to already existing techniques.

Table 3.1 Comparison of the proposed model with other models

Sr No.	Technique used	Dataset Used	Accuracy
1.	LRP, LIME Images Malolan et al. [35]	FaceForensics Faces,	94.33%
2.	CNN Agarwal et al. [31] 2020	FaceForensics++, DFDC,CELEBDF,WLDR, and DFD	98.90%
3.	CNN, KNN Frank et al. [54] 2020	GANs CelebA and LSUN	99.91%
4.	CFFN Hsu et al. [51]	Large pose variations, and background clutter	98.80%
5.	VGG-16 Qurat et al. [47]	Real and Fake Face Detection	92.09%
6.	Proposed Hybrid learning Kernel PCA based technique	Face App dataset	99.43%

3.2.1 Parameters analysis

To compare the efficiency of the proposed method with the existing work , Confusion matrix and ROC curve are analyzed to considered performance measures. Figure 3.1 and Figure 3.2 explain change in amplitude of data points and increment in sub values after dimension reduction using Kernel PCA vs original data.

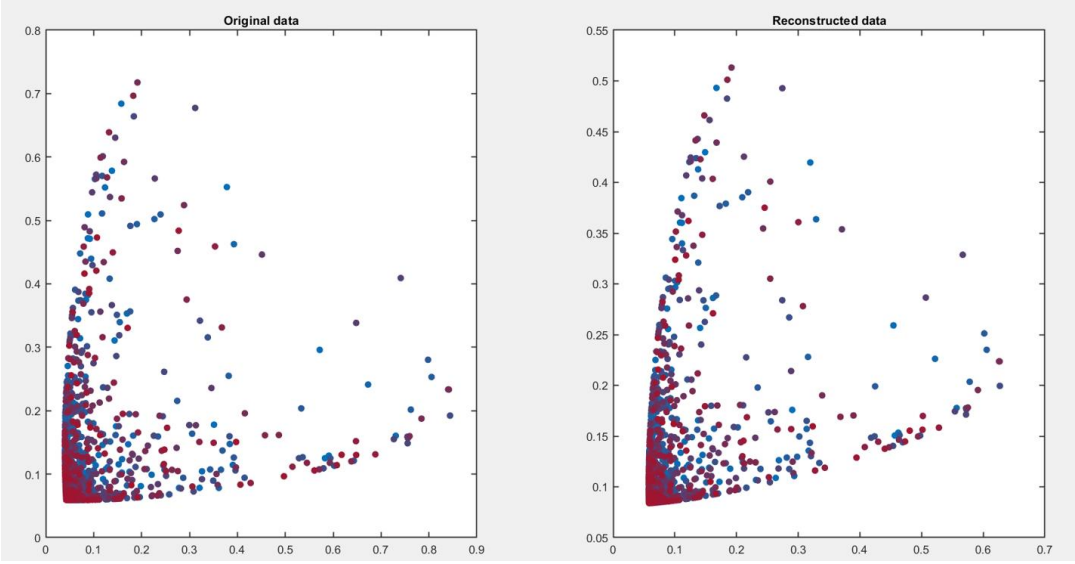

Figure 3.1 Comparison between Original data and reconstructed KPCA

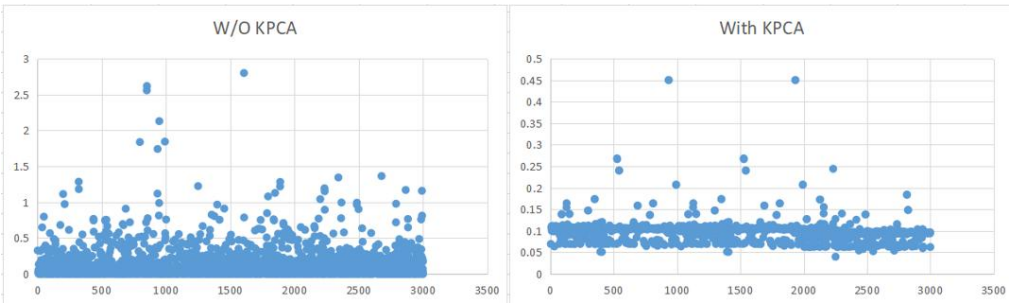

Figure 3.2 Comparison of single feature value with KPCA and without KPCA

Figure 3.3 Confusion matrix for proposed technique

Table 3.3 KPCA Model Parameters

S.No	Parameter	Value
1	Kernel function	Gaussian
2	No. of sample	998
3	No. of features	2048
4	No. of components	100
5	No. of T2 Alarm	121
6	No. of SPE alarm	0

7	Accuracy of T2	87.8758 %
8	Accuracy of SPE	100.0000%

Table 3.4 Comparison of the proposed model with different conditions

ACCbal:Balanced Accuracy, MCC: Mathews Correlation Coefficient.

Dataset	Model	ACC	ACC_{bal}	Sen %	Spec %	F score	MCC	Time (s)
3000 Indigenous images and 3000 Fake images	Medium Tree + 2048 Features	66.2	66.2	66.36	66.09	0.6610	0.3246	45.228
3000 Indigenous images and 3000 Fake images	KPCA + Coarse KNN + 2048 Features	81.3	99.52	99.11	99.93	99.52	99.04	42.62
3000 Indigenous images and 3000 Fake images	KPCA + Cubic KNN + Top 20 features	99.43	99.4	98.98	99.90	99.44	98.87	5.99
3000 Indigenous images and 3000 Fake images	NMF + Coarse KNN + 2048 Features	82.8	86.96	74.62	99.35	0.85	0.69	71.02
3000 Indigenous images and 3000 Fake images	NMF + Coarse KNN +Top 20 features	83.2	86.77	75.34	98.21	0.85	0.69	5.86

3.3 Advantages

The proposed approach effectively addresses the issue of data imbalance encountered in the Faceapp dataset. It leverages a deep neural network to extract 2048 features from the images. Furthermore, the classification into indigenous and fake categories provides a distinct understanding of the distinguishability of an image. The advantage lies in the pre-ranked features and reduced higher dimensionality,

eliminating the need for additional computational or instrumental requirements compared to other Deepfake methods. Moreover, the implementation of Kernel PCA and the prediction model demonstrates faster performance compared to conventional existing models.

3.4 Limitations

The proposed hybrid learning prediction-based model has been specifically applied to images featuring attribute manipulation, focusing solely on the impact on facial expression. However, it is important to acknowledge that deep fakes can encompass various other modalities such as audio, video, and text. The current research does not encompass parameters for measuring forgery in video or rigidity, which can be valuable additions to future studies. Additionally, future works should consider including more diverse scenarios that can affect detection, as generative networks continue to advance rapidly. Furthermore, exploring the severity of forged images can be a promising avenue for future research.

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

This study employed Kernel-based PCA and Non-Negative Matrix Factorization in combination with ResNet 50 and machine learning prediction to address the inherent imbalance in the image dataset, which caused biases in the machine learning models. To mitigate this bias, data augmentation techniques were utilized. Additionally, multiple classification models were implemented to enhance the accuracy of the predictions.

The parameters of test images, including facial swap, attribute manipulation, and gender swap, were carefully considered to enable the prediction models to better learn the distinctive characteristics of deepfake and indigenous images. These features play a crucial role in the manifestation and detection of deepfake instances.

Moving forward, future research aims to improve the accuracy of the current methodology and enhance pre-processing techniques to enhance image quality. While the current dataset comprises high-quality input images, it is important to develop robust pre-processing techniques to improve the quality of the results, ensuring the system is capable of accurately classifying indigenous and forged images. Furthermore, in addition to accuracy, incorporating calculations such as precision, F1 score, and Matthew's Correlation Coefficient (MCC) provides more comprehensive information about the model's performance, as they consider the negative predicted class in their calculations.

REFERENCES

[1] Westerlund, M. 2019. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11): 40-53. http://doi.org/10.22215/timreview/1282

[2] Kaliyar, R.K., Goswami, A. & Narang, P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. *J Supercomput* **77**, 1015–1037 (2021). https://doi.org/10.1007/s11227-020-03294-y

[3] Guarnera, L., Giudice, O., Nastasi, C., & Battiato, S. (2020, September). Preliminary forensics analysis of deepfake images. In *2020 AEIT international annual conference (AEIT)* (pp. 1-6). IEEE.

[4] Lv, Z., Zhang, S., Tang, K., & Hu, P. (2022, May). Fake audio detection based on unsupervised pretraining models. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 9231-9235). IEEE.

[5] Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology innovation management review, 9(11).

[6] Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights, 1(1), 100004.

[7] Mbona, I., & Eloff, J. H. (2022). Feature selection using Benford's law to support detection of malicious social media bots. *Information Sciences*, *582*, 369-381.

[8] Chen, Y., Yang, X. H., Wei, Z., Heidari, A. A., Zheng, N., Li, Z., ... & Guan, Q. (2022). Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 105382.

[9] Zhang, T. (2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5), 6259-6276.

[10] Swathi, P., & Sk, S. (2021, September). Deepfake creation and detection: A survey. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 584-588). IEEE.

[11] Lyu, S. (2020, July). Deepfake detection: Current challenges and next steps. In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)* (pp. 1-6). IEEE.

[12] T.-T. Ng, J. Hsu, and S.-F. Chang. Columbia image splicing detection evaluation dataset, 2009. 2, 5

[13] J. Dong, W. Wang, and T. Tan. Casia image tampering detection evaluation database 2010. In http://forensics.idealtest.org. 2, 5

[14] J. Dong,W.Wang, and T. Tan. Casia image tampering detection evaluation database. In Signal and Information Processing (ChinaSIP), 2013 IEEE China

Summit & International Conference on, pages 422–426. IEEE, 2013. 2, 5

[15] T. J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha. Exposing digital image forgeries by illumination color classification. ieee transactions on information forensics and security, 8(7):1182– 1194, 2013. 1,2, 3, 5

[16] Theckedath, D., & Sedamkar, R. R. (2020). Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science*, *1*, 1-7.

[17] The Annotated ResNet-50 Mukherjee https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[18] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition* (pp. 5781-5790).