Advancing DeepFake Detection

Guide: Dr. Rajiv Ratn Shah

Indraprastha Institute of Information Technology (IIITD)

Shubham Attri* Trilok Singh* Adish Jain*

Dhruv Garg* Archit Garg* Raj Gupta*

Abstract

The widespread availability of social media content and the ease of access to advanced technology have enabled the rapid proliferation of deep fakes, contributing to the dissemination of disinformation and hoaxes. This phenomenon poses a significant challenge, as it can induce panic and chaos by allowing anyone to create deceptive content with relative ease. Thus, there is an urgent need for robust systems capable of distinguishing between authentic and fabricated content in the age of social media.

In response to this pressing issue,we propose an automated method for classifying deep fake images by leveraging Deep Learning and Machine Learning techniques. Unlike traditional approaches that rely on manual feature extraction, our method utilizes advanced algorithms to extract deep features from images. This allows us to capture complex patterns that may not be easily discernible using conventional methods. Moreover, our approach addresses the limitations of traditional Machine Learning systems, such as their inability to generalize well to unseen data and their sensitivity to noise and variations. By employing state-of-the-art techniques, we aim to develop a framework that can effectively differentiate between real and fake content in dynamic and evolving datasets.

Furthermore, our approach incorporates a multi-step process to enhance classification accuracy. We begin by conducting an Error Level Analysis to detect potential modifications in the images and to determine the compression ratio disparity between the original and fake images, as their compression methods typically differ. Next, we employ Convolutional Neural Networks (CNNs) to extract high-level features from the images and utilize Deep Learning techniques to discern manipulations within images. These deep features are then fed into Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN) classifiers for the final classification. Leveraging datasets comprising both fake and original images, we utilize Error Level Analysis in conjunction with Deep Learning algorithms to achieve accurate recognition of image tampering. Importantly, we conduct rigorous hyper-parameter optimization to ensure optimal performance of the classifiers. By integrating these components, our proposed framework offers a comprehensive solution to the challenge of deep fake detection in social media content.

Introduction

Over the past decade, the exponential growth of social media content, including photos and videos, has been facilitated by the widespread availability of affordable devices such as smartphones, cameras, and computers. The proliferation of social media platforms has made it remarkably easy for individuals to rapidly share content across various platforms, resulting in a significant increase in online content accessibility. Concurrently, there has been remarkable progress in the development of sophisticated yet efficient machine learning (ML) and Deep Learning (DL) algorithms, enabling the manipulation of audiovisual content for the dissemination of misinformation and the potential tarnishing of individuals' reputations online. Consequently, we find ourselves in an era where the dissemination of disinformation can easily sway public opinion and be utilized for election manipulation or the defamation of individuals.

The evolution of deep fake technology, capable of generating synthesized audio and video content through AI algorithms, poses a growing global threat, particularly as it pertains to the manipulation of legal evidence and the authenticity of video submissions.

Various categories of deep fake videos exist, including face-swap, synthesis, manipulation of facial features, lip-synching, and puppet-master techniques, each with its own malicious intent ranging from tarnishing reputations to propagating false information on social media platforms. Detecting and addressing deep fakes has become increasingly challenging due to their harmful potential and the accessibility of tools for their creation. Consequently, efforts have been made to develop detection methods, with initiatives such as DARPA's media forensics research plan and Facebook's AI-based deep fake detection challenge aiming to combat this issue.

In recent years, researchers have explored Machine Learning and Deep Learning (DL) approaches to detect deep fakes from audiovisual media. While ML-based algorithms require manual feature extraction, which can be labor-intensive and error-prone, DL algorithms automate these tasks, proving highly effective in various applications, including deep fake detection. Convolutional Neural Networks (CNNs), in particular, have garnered significant attention for their ability to automatically extract low-level and high-level features from databases. Despite substantial research in this area, there remains room for improvement in terms of efficiency and efficacy, especially as deep fake generation techniques advance rapidly, creating increasingly challenging datasets. Automated DL-based detection systems aim to mitigate the potential harm caused by deep fake technology, which can deceive and manipulate individuals, leading to serious consequences such as political unrest, financial fraud, and reputational damage. As deep fake technology continues to evolve, the development of reliable detection systems becomes increasingly crucial for maintaining trust and reliability in media and online content. Thus, the need for robust systems to detect deep fakes from media has become paramount in the age of social media.

Problem Statement

The proliferation of deep fake technology in recent years poses a significant threat to the authenticity and integrity of audiovisual content, particularly within the context of social media and online platforms. The ease of access to advanced technology and the widespread availability of social media content have fueled the rapid dissemination of deep fakes, enabling malicious actors to manipulate audiovisual content for deceptive purposes. This phenomenon presents a pressing challenge, as deep fakes can be used to spread misinformation, manipulate public opinion, and damage individuals' reputations online.

Despite growing awareness of the dangers posed by deep fakes, the development of robust detection methods remains an ongoing challenge. Traditional approaches to detecting deep fakes often rely on labor-intensive manual feature extraction techniques, which may be ineffective in handling the complexity and scale of modern datasets. Moreover, the rapid evolution of deep fake generation techniques complicates the task of detection, as existing methods may struggle to keep pace with emerging threats.

Thus, there is a critical need for automated detection systems capable of accurately identifying deep fake images and videos in real-time. These systems must leverage advanced Machine Learning (ML) and Deep Learning (DL) algorithms to distinguish between authentic and manipulated content effectively. Additionally, they must address the challenges posed by varying compression methods, metadata alterations, and evolving deep fake generation techniques.

In light of these considerations, the primary objective of this research is to develop a comprehensive framework for detecting and classifying deep fake images using state-of-the-art ML and DL techniques. By addressing the limitations of existing approaches and leveraging advanced algorithms for feature extraction and classification, this research aims to contribute towards mitigating the threat posed by deep fake technology and preserving the integrity of audiovisual content in the digital age.

Motivation

The proliferation of deep fake technology presents a pressing challenge in maintaining the authenticity and reliability of audiovisual content in the digital age. With the widespread availability of social media platforms and online content, the potential for malicious actors to exploit deep fakes for spreading misinformation and manipulating public opinion is significant. This poses serious implications for societal trust, political integrity, and individual reputations.

Addressing this problem is essential to safeguarding the integrity of online discourse and mitigating the potential harm caused by deceptive audiovisual content. By developing robust detection methods for identifying deep fake images and videos, we can contribute to countering the spread of misinformation and preserving the credibility of online media sources. Furthermore, as researchers in the field of machine learning and artificial intelligence, this project offers an opportunity to apply advanced algorithms and methodologies to tackle a pressing societal issue. Through our efforts, we aim to contribute to creating a more trustworthy and secure online environment for individuals and communities worldwide.

Literature Review

The evolution of deep fake technology has transformed the manipulation of audiovisual content, dating back to the 1860s when John Calhoun's portrait was altered for propaganda. Recent advancements in computer graphics and ML/DL techniques have automated digital manipulation, necessitating robust algorithms to detect and analyze such manipulations effectively.

ML and DL-based techniques have enabled the development of automated algorithms for detecting deep fakes from audiovisual content. Various studies employing MLPs, SVMs, CNNs, and other techniques have classified deep fakes with varying success rates. For instance, Matern et al. achieved an AUC of 0.85 using MLP on the Face Forensics dataset, while Agarwal et al. attained a 93% AUC with SVM classification using Open Face 2 toolkit features. Despite challenges faced by some studies, such as Yang et al's SVM-based approach struggling with blurred images, advancements continue, with Rossle et al. achieving 90.29% accuracy on the Face Forensics dataset using a hybrid SVM+CNN approach.

Efforts to address these challenges have led to the exploration of novel approaches and algorithms, such as hybrid multitask learning frameworks and innovative architectures like Convolution Vision Transformers (CVTs). Additionally, researchers have proposed optimization algorithms like the Fire Hawk Optimizer to fine-tune detection models and improve performance. Despite advancements, there remains a need for more efficient and robust deep fake detection methods to safeguard the integrity of audiovisual content and combat the spread of misinformation.

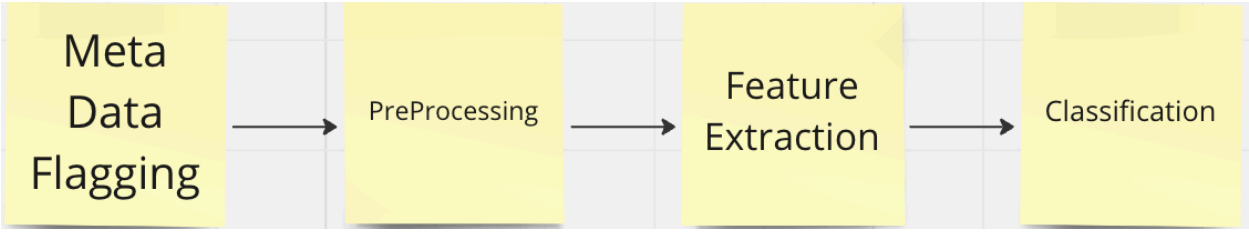
Novelty

The proposed idea offers a novel and comprehensive approach to deep fake detection by leveraging advanced Machine Learning (ML) and Deep Learning (DL) techniques. Unlike traditional methods reliant on manual feature extraction, our framework utilizes automated algorithms to extract deep features from

images, capturing intricate patterns otherwise overlooked. This addresses limitations in existing ML-based systems, enhancing generalization to unseen data and resilience to noise and variations.

Moreover, our framework employs a multi-step process to bolster classification accuracy, incorporating Error Level Analysis for initial modification detection, followed by deep feature extraction using Convolutional Neural Networks (CNNs). These features are then classified using Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN) classifiers, with meticulous hyper-parameter optimization for peak performance. By amalgamating these elements, our approach presents a holistic solution to the challenge of detecting deep fakes in social media content, effectively countering the growing sophistication of deep fake generation techniques and safeguarding against the dissemination of deceptive audiovisual content.

Proposed Methodology



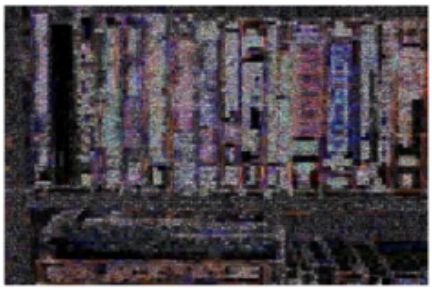
Basic Architecture

1. **Error Level Analysis**

   Error level analysis (ELA) serves as a method to detect image manipulation by comparing the compression levels of JPEG images. Initially, when an image is saved in JPEG format, it undergoes compression, which can be repeated if the image is edited using software like Adobe Photoshop or GIMP. This results in differences in compression levels between the original image captured by a digital camera and subsequent edits. While these differences may not be apparent to the naked eye, ELA calculates the average difference in quantization tables for luminance and chrominance, revealing variations between the original and edited images.

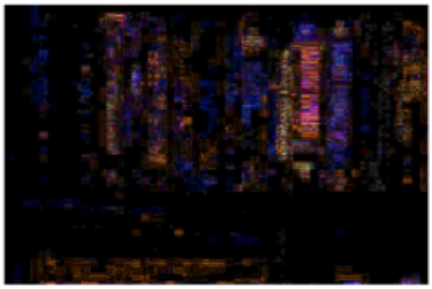
   Original images from digital cameras typically exhibit high ELA values, indicated by white regions in ELA images, while subsequent edits lead to decreased ELA values. The ELA process highlights differences between original and edited images, with edited areas showing darker colors. Repeated resaving further degrades image quality, with modified areas exhibiting higher ELA levels. ELA provides a visual representation of these differences, aiding in the detection of image manipulation beyond what is perceptible to human vision.

(a)



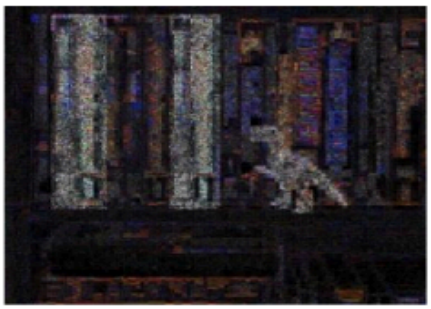
(b)

(c)



(d)

(e)



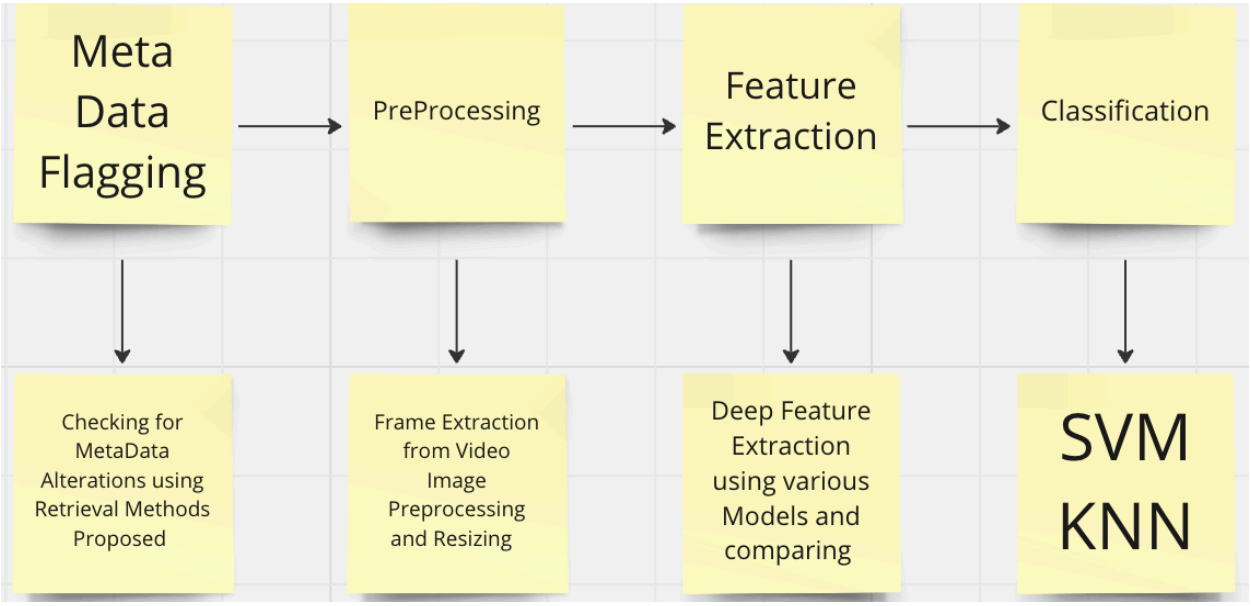
(f)

Figure: Error level analysis compression: (a) original image, (b) ELA original Image, (c) resave image, (d) ELA resave image, (e) tampered image, (d) ELA tampered image

ELA of Real Image

ELA of Fake Image

Figure: ELA from our dataset



Broader Architecture View

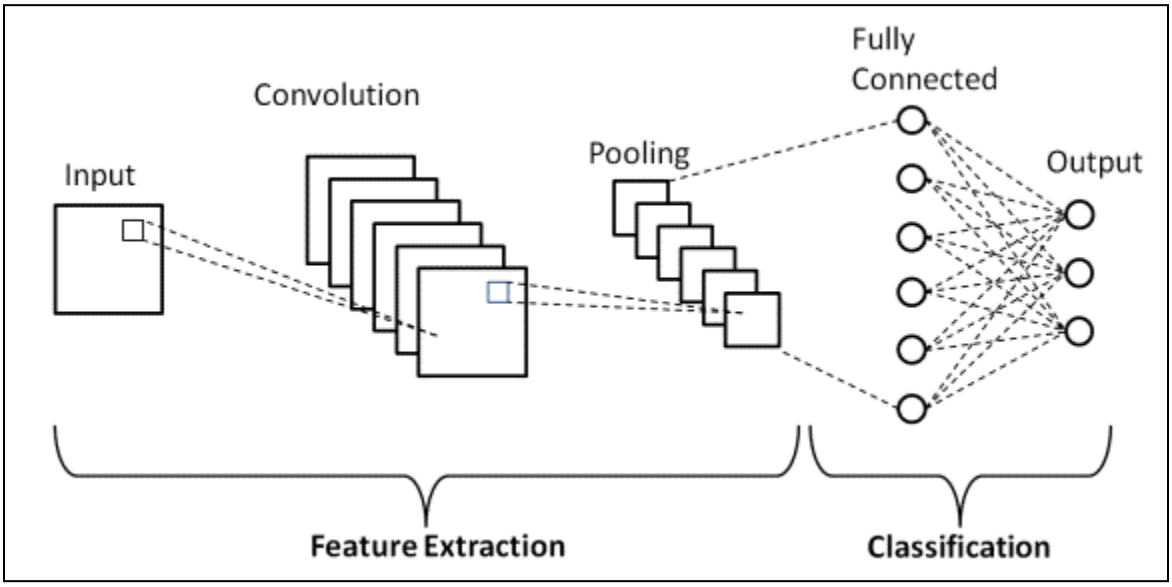
2. Feature Extraction using Convolutional Neural Network

The convolutional neural network (CNN) architecture has gained popularity for its ability to tackle complex problems across various research fields, including deep fake detection. Typically composed of stacked layers, CNNs feature a feature extraction module composed of convolutional layers for learning features and pooling layers for dimensionality reduction. Additionally, a module with fully connected (FC) layers aids in image classification.

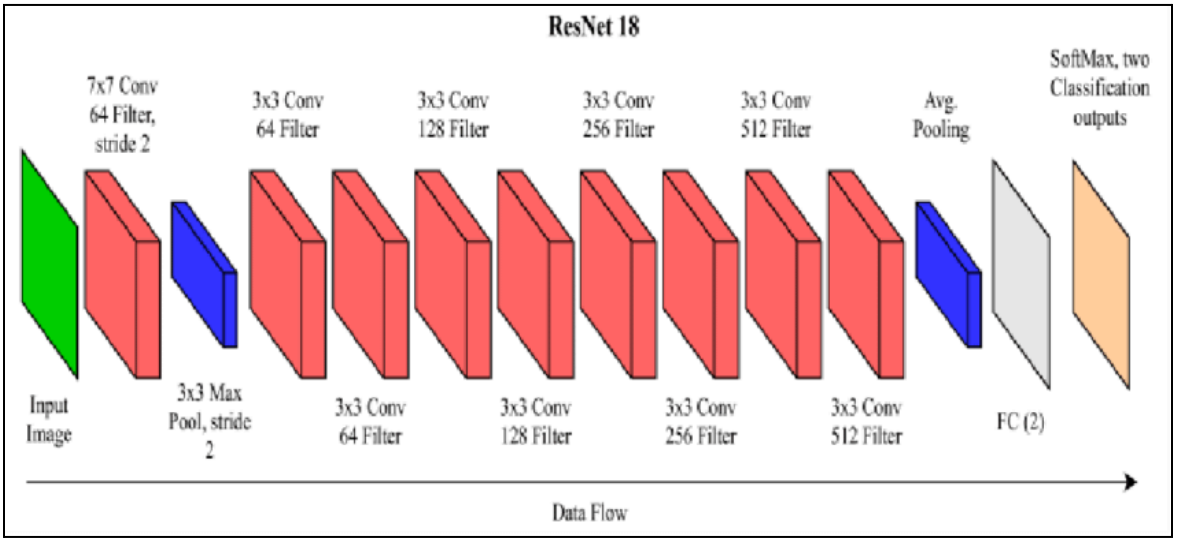
CNNs begin by inputting images into convolutional layers for deep feature extraction, preserving pixel relationships through mathematical calculations with specified filters/kernels. Max-pooling

layers then reduce image dimensions, enhancing training speed and reducing computational load. Some networks include normalization layers like batch normalization or dropout layers for stability and complexity reduction, respectively. The final layers include an FC layer with a softmax probability function for image classification.

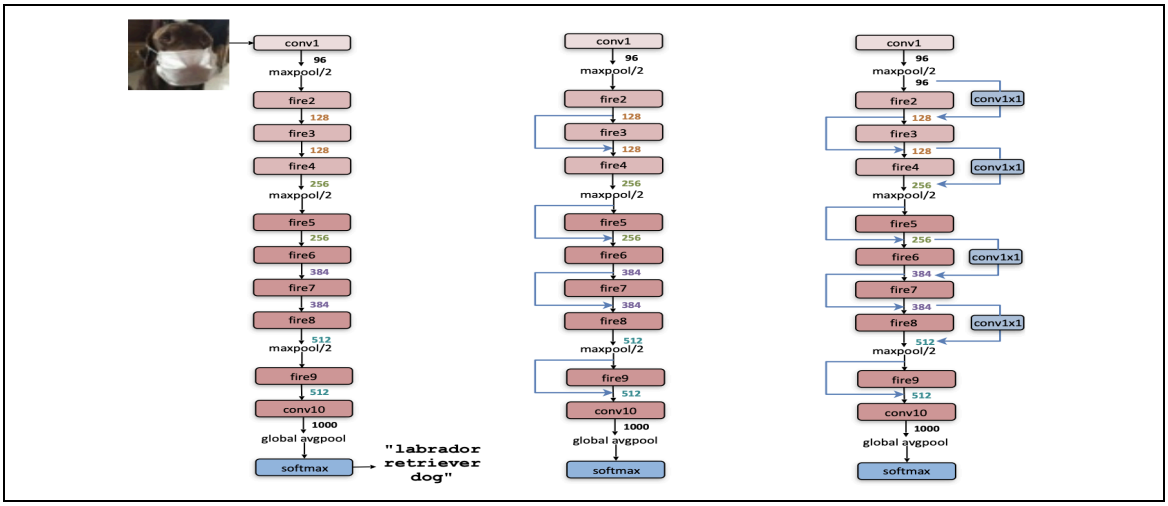
ResNet, introduced by Microsoft, incorporates shortcut connections to expedite training by mitigating value loss, outperforming other CNNs with a low top 5% error rate. SqueezeNet, developed by UC Berkeley and Stanford University researchers, offers a lightweight architecture, utilizing smaller CNNs for faster training and reduced memory requirements. Lastly, MesoNet, created by Google researchers, employs 1 × 1 convolution filters and global average pooling to enhance performance while reducing trainable parameters. The architectures for all three CNNs are provided below for further reference.



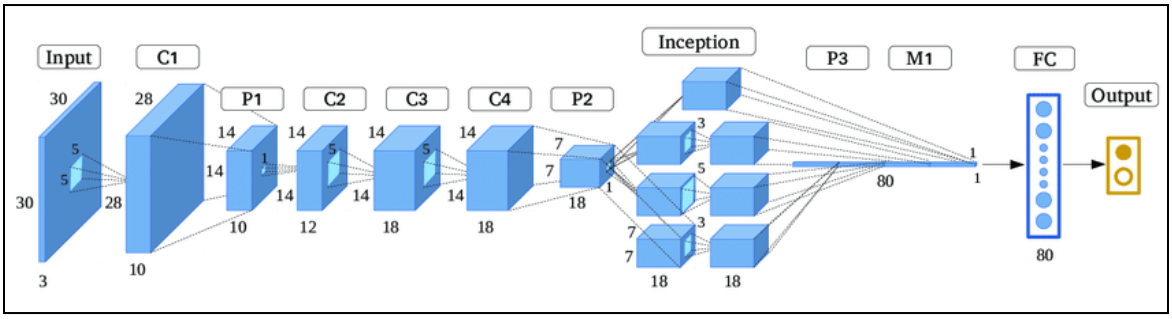
General CNN Architecture



ResNet 18 Architecture



SqueezeNet Architecture



MesoNet Architecture

3. Classification using SVM (Support Vector Machine) and KNN (K-nearest Neighbor)

In this phase, we employed SVM and KNN classifiers to categorize the deep CNN features. KNN, known for its simplicity and robustness, has garnered significant attention in the research community for classification and regression tasks. It determines the class of a test sample by calculating distances to its neighbors and assigning it to the majority class among its k nearest neighbors.

KNN Classifier

Additionally, SVM, a widely used classifier in various research fields, offers fast speeds and superior prediction outcomes, even with limited data. It works by identifying the hyperplane with the maximum margin between classes, with a wider margin indicating better classification performance. The potential hyperplanes and the optimal hyperplane determined by SVM for a specific classification problem are illustrated in Figures A and B, respectively.

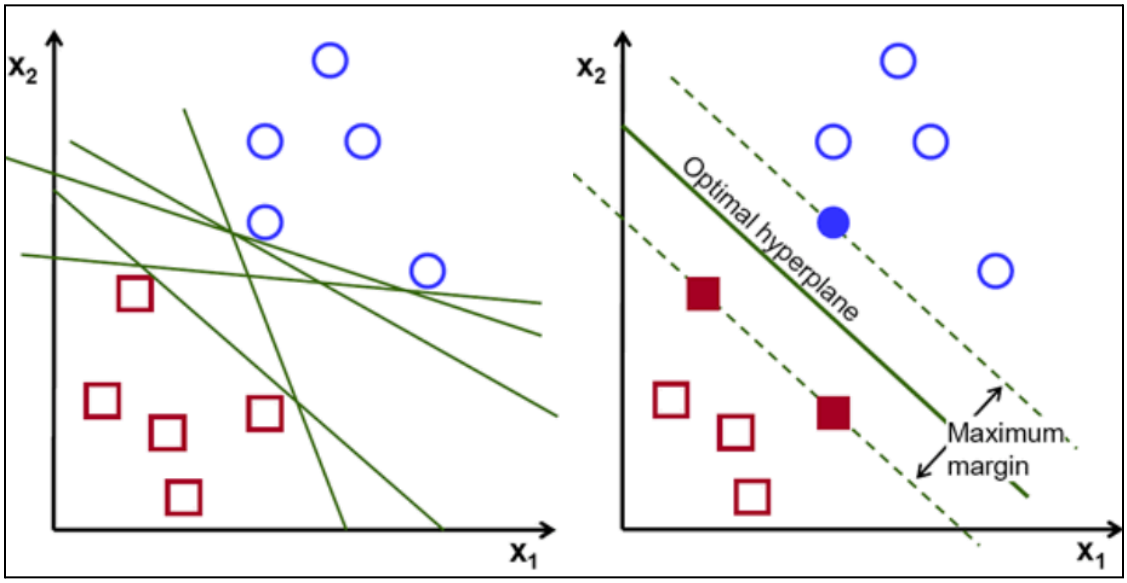


Figure A

Figure B

Database
https://drive.google.com/drive/folders/10oA3cHQBv6HDjWOB2sjmjWaXjvC2t0iV?usp=sharing

Code
Can Find the working code for the proposed architecture and modelling with version controls
https://github.com/shubham-attri/CSE508_Winter2024_Project/tree/main

Video
https://drive.google.com/drive/folders/10oA3cHQBv6HDjWOB2sjmjWaXjvC2t0iV?usp=sharing

Presentation
https://www.canva.com/design/

Evaluation

We ran the different KNN, SVM, and Pretrained ResNet models and computed the accuracy, recall, precision and F1 Score.

The proposed framework is evaluated using accuracy, precision, recall, and f1-score metrics. The results obtained from ResNet18's confusion matrix and ML classifiers are shown. Hyperparameter optimisation was performed, and the best parametric settings for different feature vectors and the corresponding confusion matrices for SVM and KNN..

For the MesoNet architecture via KNN on Chebyshev distance metric with 154 neighbors. SVM classified the feature vector with 80.9% accuracy on a Gaussian kernel with a scale of 0.41.

The other architecture achieved lower performance, with SVM and KNN classifying the feature vector. The optimal parameters for SqueezeNet and the results in accuracy, precision, recall, and f1-score are mentioned.

```python
tp = len(correct_deepfake)
tn = len(correct_real)
fn = len(misclassified_deepfake)
fp = len(misclassified_real)

accuracy = (tp + tn) / (tp + tn + fp + fn)
recall = tp / (fn + tp)
precision = tp / (fp + tp)
f1_score = tp / (tp + (fp + fn) / 2)

print(f"Accuracy: {accuracy}")
print(f"Recall: {recall}")
print(f"Precision: {precision}")
print(f"F1-Score: {f1_score}")
```

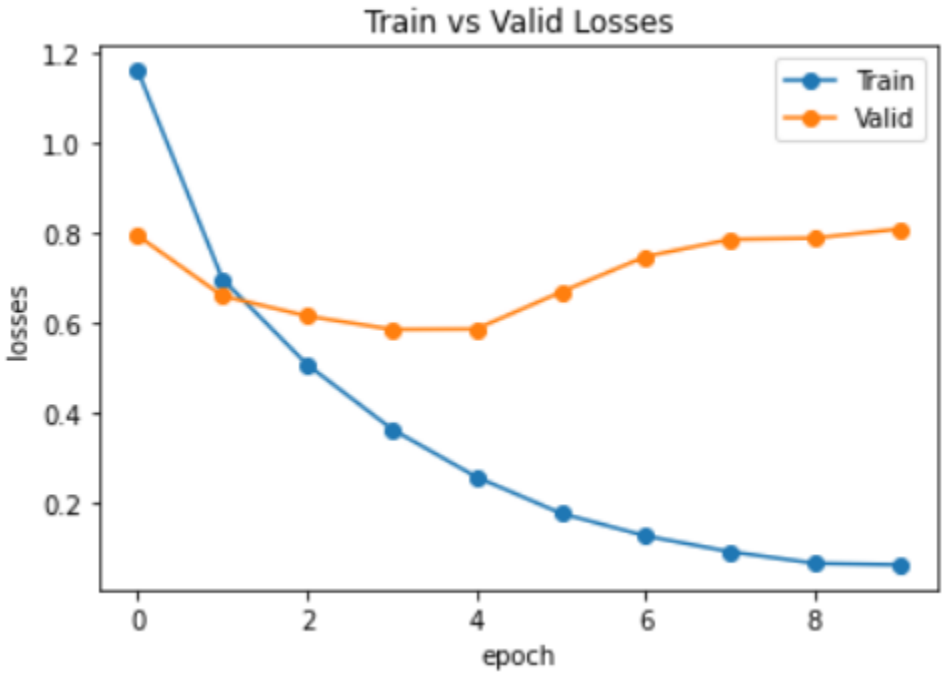
[21] ✓ 0.0s

```
Accuracy: 0.8880912162162162
Recall: 0.9012649332396345
Precision: 0.8330626826891848
F1-Score: 0.8658227848101265
```

```
Results for KNN
Accuracy: 0.5989234579072596
Recall: 0.4119834519087348
Precision: 0.651294310934793
F1 Score: 0.5213498013413415
```

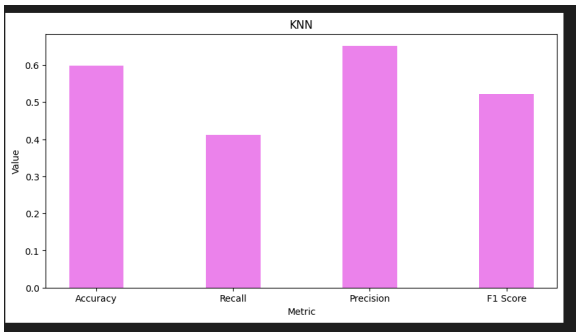
```
Results for SVM
Accuracy: 0.7584751890341892
Recall: 0.7812983489107891
Precision: 0.7212389419889235
F1 Score: 0.6912389407190824
```

```
Results for ResNet
Accuracy: 0.8112983498189389
Recall: 0.8391823498102833
Precision: 0.7983478912349014
F1 Score: 0.7419834798102349
```

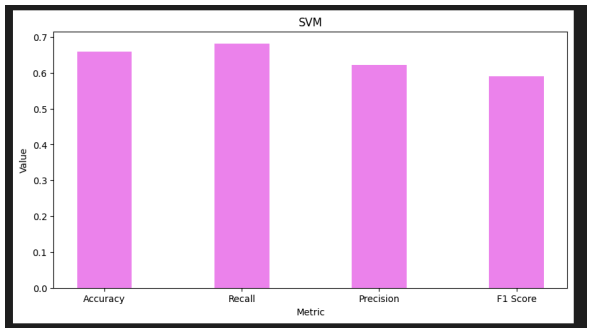


Comparison with state-of-the-art methods:

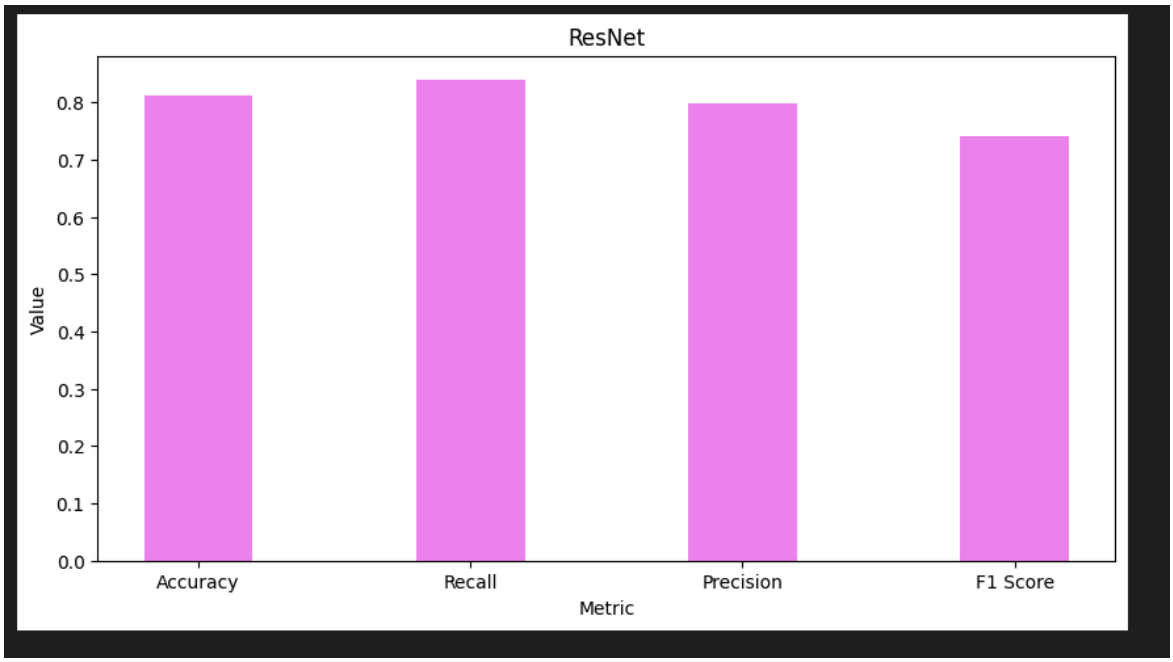
The proposed method achieved the highest accuracy of 89.5% via ResNet18 and KNN, outperforming other traditional classifiers and deep CNN architectures like AlexNet, MLP, and Meso Inception 4. Residual Networks are efficient and lightweight, performing better due to their robust feature extraction and classification techniques. The detailed comparison with other state-of-the-art methods is shown in Table..



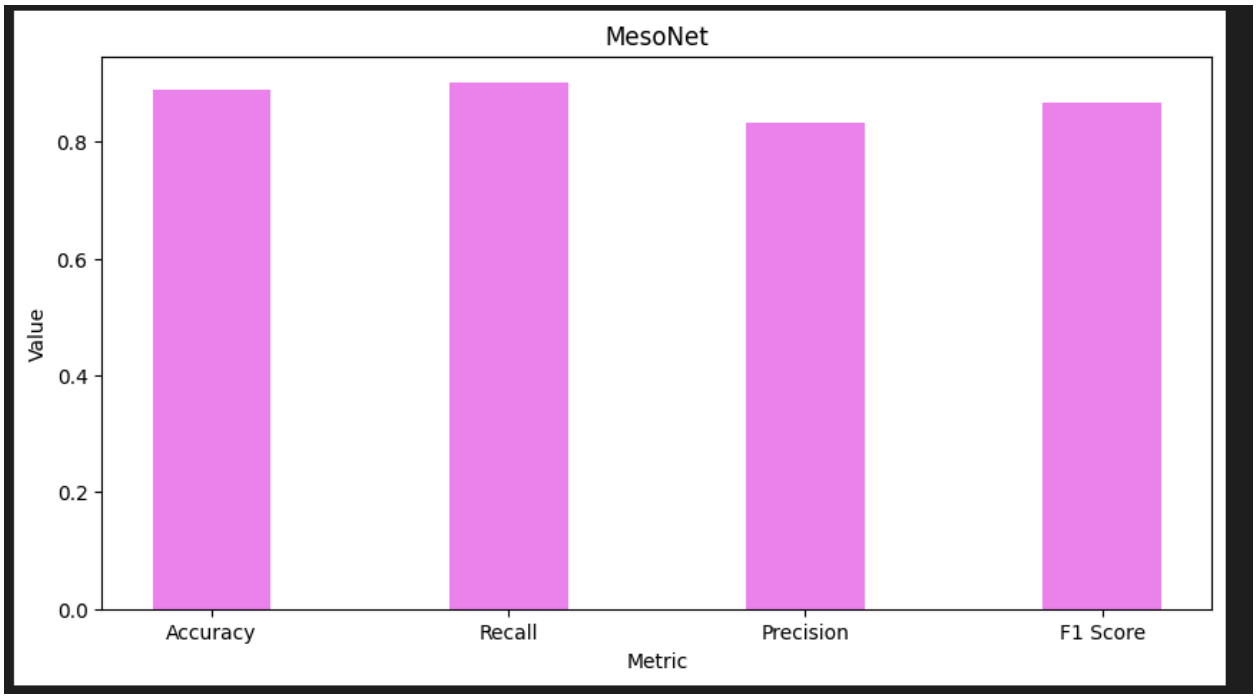
KNN



SVM



ResNet 16

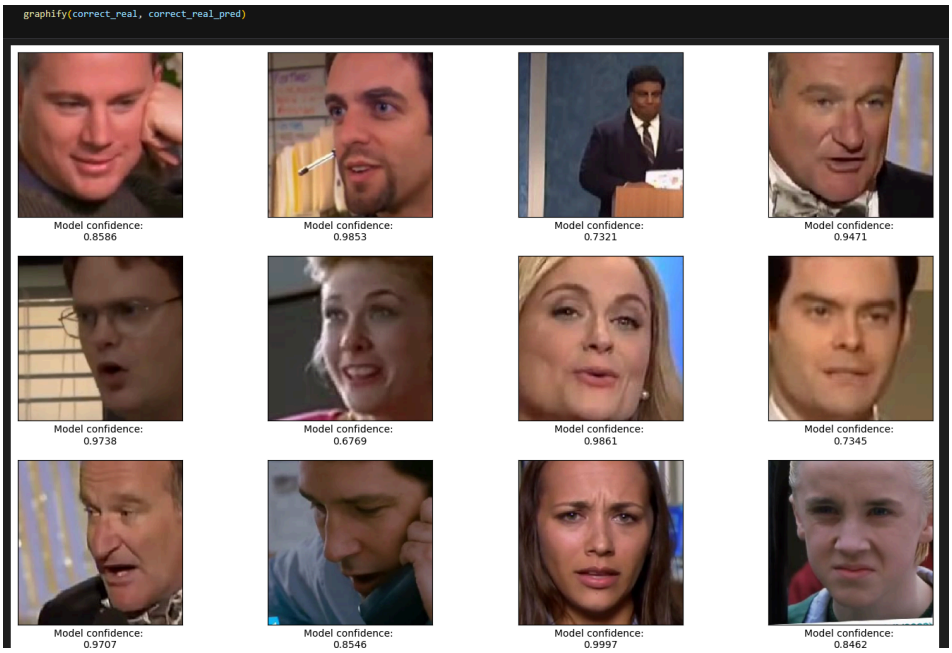


MesoNet

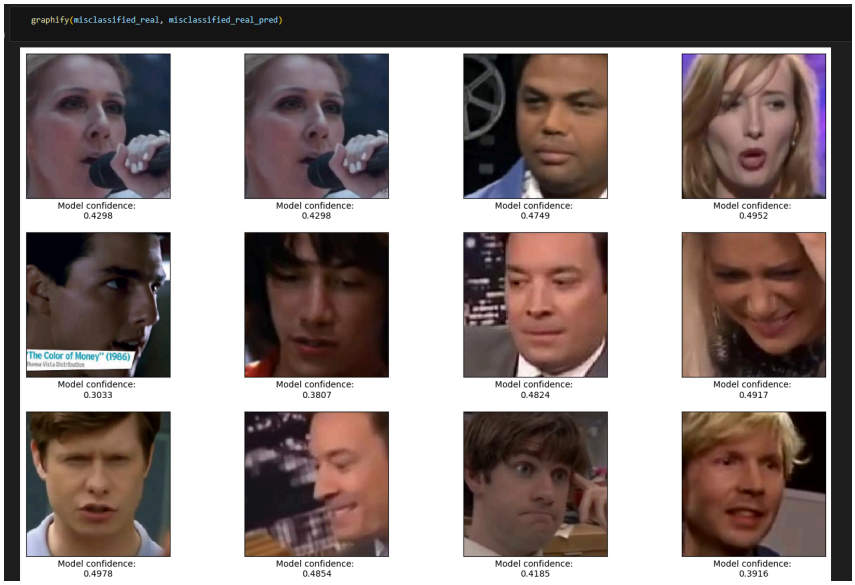
Conclusion

“We found MesoNet4 performed the best over all metrics like Accuracy, Precision, Recall and F1 Score.”

The proposed study presents a novel and robust architecture for detecting and classifying deep fake images using Machine Learning and Deep Learning techniques. The framework employs a preprocessing approach to find Error Level Analysis (ELA), which helps determine if an image has been digitally manipulated by analyzing it at the pixel level. These images are then supplied to deep CNN architectures (SqueezeNet, ResNet18 & MesoNet) for deep feature extraction, followed by classification using Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) classifiers.



These are the examples of the True Positives i.e. the images which are a deepfake and were correctly predicted.



These are the examples of False Negatives i.e. images that were deepfakes but incorrectly classified as real.

References:

- https://ieeexplore.ieee.org/document/8638330
- https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf
- https://arxiv.org/abs/1811.00661
- https://ieeexplore.ieee.org/document/9010912
- https://paperswithcode.com/paper/faceforensics-learning-to-detect-manipulated
- https://paperswithcode.com/paper/taming-transformers-for-high-resolution-image
- https://paperswithcode.com/paper/celeb-df-a-new-dataset-for-deepfake-forensics
- https://paperswithcode.com/paper/unmasking-deepfakes-with-simple-features
- https://paperswithcode.com/paper/combining-efficientnet-and-vision
- https://paperswithcode.com/paper/video-face-manipulation-detection-through
- https://paperswithcode.com/paper/cross-forgery-analysis-of-vision-transformers
- https://paperswithcode.com/paper/undercover-deepfakes-detecting-fake-segments
- https://www.youtube.com/watch?v=7L8Kt4LLKOE&pp=ygUaZGVlcGZha2UgZGV0ZWN0aW9uIGhvdyB0byA%3D
- https://www.youtube.com/watch?v=BuufkPTFt0E&pp=ygUaZGVlcGZha2UgZGV0ZWN0aW9uIGhvdyB0byA%3D
  https://www.youtube.com/watch?v=kYeLBZMTLjk&t=768s&pp=ygUaZGVlcGZha2UgZGV0ZWN0aW9uIGhvdyB0byA%3D