“Understanding Real and Fake Faces"

A project report submitted in partial fulfillment of the requirements for the DSE 315/615 course at Indian Institute of Science Education and Research Bhopal

by

Agniva Banerjee (2410701)
EECS Department

Hiba K T (22146)
DSE Department

Course Instructor: Dr. Samiran Das
Data Science & Engineering (DSE) Department

Indian Institute of Science Education and Research Bhopal
Bhopal, Madhya Pradesh, India
Date: 29.11.2024

Contents

List of Figures

List of Tables

Abstract

Title: Understanding Real and Fake Faces

Agniva Banerjee (2410701), Hiba K T (22146)

A project report submitted in partial fulfillment of the requirements for the course DSE 315/615

Indian Institute of Science Education and Research Bhopal
Bhopal

Course Instructor: Dr. Samiran Das
Data Science & Engineering (DSE) Department, IISER Bhopal

In the campaign against synthetic media, ennobling fake face detection is influential. Here we address the challenge of face image classification by leveraging dimensionality reduction techniques to manage the high-dimensional nature of the dataset. The dataset, consisting of real and fake face images, presents a significant computational burden due to its high dimensionality. We employ Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to reduce the dimensionality of the images, aiming to enhance computational efficiency and classification performance. Finally, this study utilizes feature maps in conjunction with explainable artificial intelligence (AI) methodologies, notably SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), to execute a remarkable detection algorithm. Feature maps are analyzed, to identify significant visual patterns that are affecting accurate and inaccurate classifications. This dual approach deepens the understanding of model constraints and guides strategies for increasing robustness and accuracy in fake face detection systems. The results highlight the significance of explainable AI in creating reliable solutions for practical applications. These improvements provide considerable concerns in multiple areas, including security, privacy, and misleading information.

Chapter 1

Introduction

The rapid advancement of AI technology has led to the creation of highly realistic synthetic faces, known as deepfakes [1]. As the spread of fake media increases, the need for effective detection systems has become more critical [2]. Traditional detection methods often rely on classification algorithms to differentiate between real and synthetic images [3, 4, 5]. However, these approaches lack clarity in explaining why errors occur, making it difficult to understand the reasons behind misclassifications [6, 7]. To address this issue, this study begins by applying dimensionality reduction techniques like PCA and LDA on the dataset. PCA reduces dimensions by capturing the maximum variance in the data, while LDA focuses on maximizing class separation. Scatterplots of PCA and LDA outputs reveal key differences: PCA highlights overall variance, while LDA prioritizes class distinction, which can improve classification accuracy. These techniques simplify high-dimensional image data and provide insights for better face detection and classification.

A custom convolutional neural network (CNN) architecture [Algorithm 1] is developed to generate feature maps, followed by a classification process using LazyPredict[8]. SHAP [9] and LIME [10] are applied to explain the model's decisions and enhance interpretability. These explainable AI techniques help identify the factors leading to correct or incorrect predictions, improving the transparency and accuracy of the detection system. Combining a custom CNN model, and explainable AI methods emphasizes the importance of transparency in AI systems, building trust, and enabling real-world applications to combat misinformation [11]. By integrating these approaches, this study contributes to the development of reliable and ethical AI systems for detecting deepfakes and addressing related challenges

1.1 Dataset Description

The Real and Fake Face Detection dataset [12] on Kaggle is designed to facilitate the development of machine learning (ML) models for detecting deepfake faces. It consists of 1081 real and 960 fake images—easy, medium, and hard to detect. All images are in RGB format (3 channels) and have high dimensionality. This balanced dataset is ideal for training and testing deepfake detection models, with challenges increasing across fake difficulty levels.

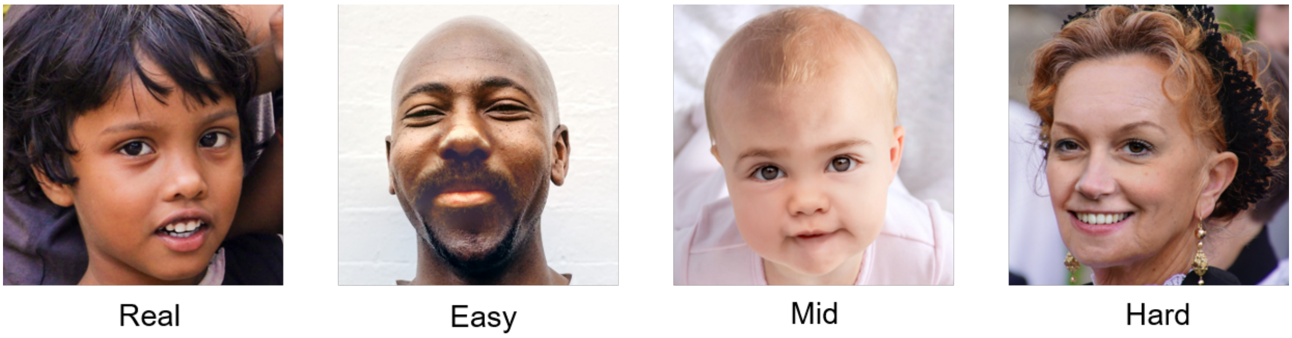

Figure 1.1: Dataset containing real and three kinds of fake images (easy, mid, hard).

1.2 PCA

PCA is a dimensionality reduction technique that simplifies high-dimensional datasets by identifying principal components, and the directions of maximum variance in the data. By projecting data onto these components, PCA reduces dimensions, enhances computational efficiency, and aids visualization. As an unsupervised method, it is useful for exploratory analysis. The process includes standardizing data, computing the covariance matrix, and projecting onto top eigenvectors. PCA is widely used in ML, 2D/3D visualization, and noise reduction while preserving essential data patterns.

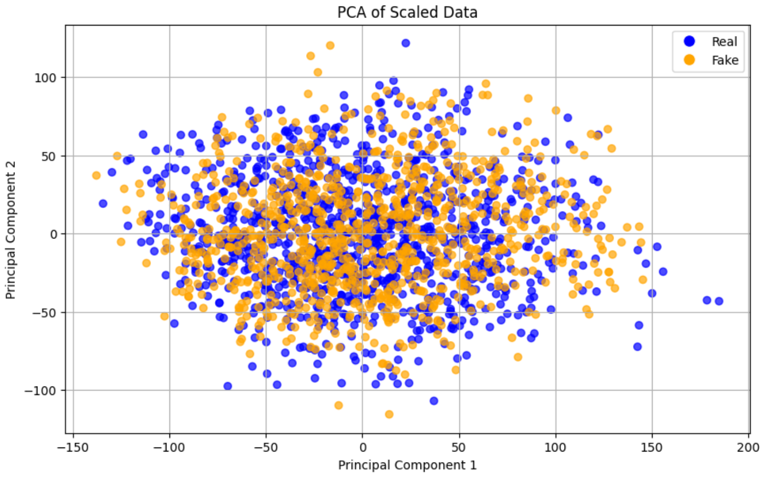

Figure 1.2: Scatter Plot of the dataset after applying PCA.

1.3 LDA

LDA is a dimensionality reduction technique focused on maximizing class separability in labeled datasets. It projects data onto a lower-dimensional space where the separation between different classes is maximized. Unlike PCA, LDA considers class labels, making it a supervised method. The process involves computing the mean and scatter matrices for each class, finding the eigenvectors of the matrix ratio, and projecting data onto the most discriminative axes. LDA is commonly used for classification tasks, feature extraction, and improving model performance. It simplifies data while emphasizing inter-class differences.

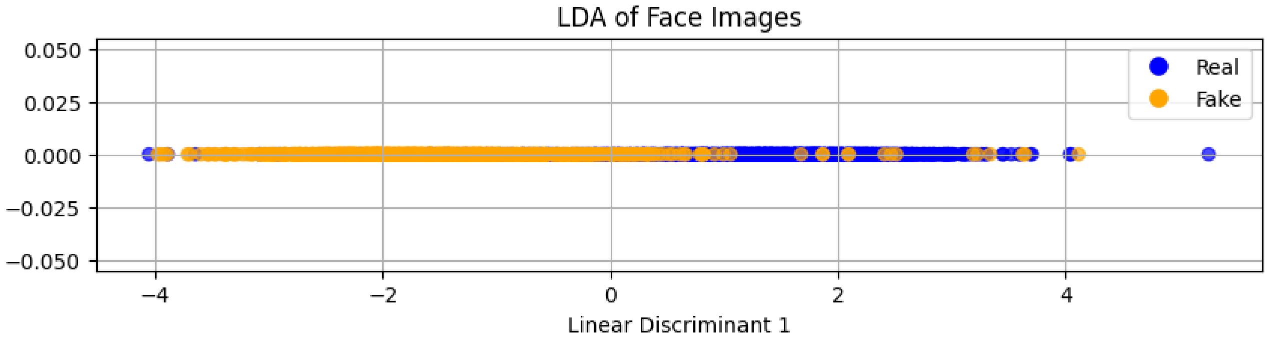

Figure 1.3: Dataset Plotted after applying LDA.

1.4 Eigen Face

Eigenfaces is a facial recognition technique that uses PCA to reduce the dimensionality of facial image data while retaining the most significant features. The process involves collecting facial images, centering them by subtracting the mean image and calculating a covariance matrix to capture the variations in the data. The eigenvectors of this matrix, known as eigenfaces, represent the key variations in the dataset. These eigenfaces are then used to project new face images into a lower-dimensional space for recognition. By focusing on the most important features, eigenfaces simplify face recognition tasks, making them computationally efficient and effective for large datasets.

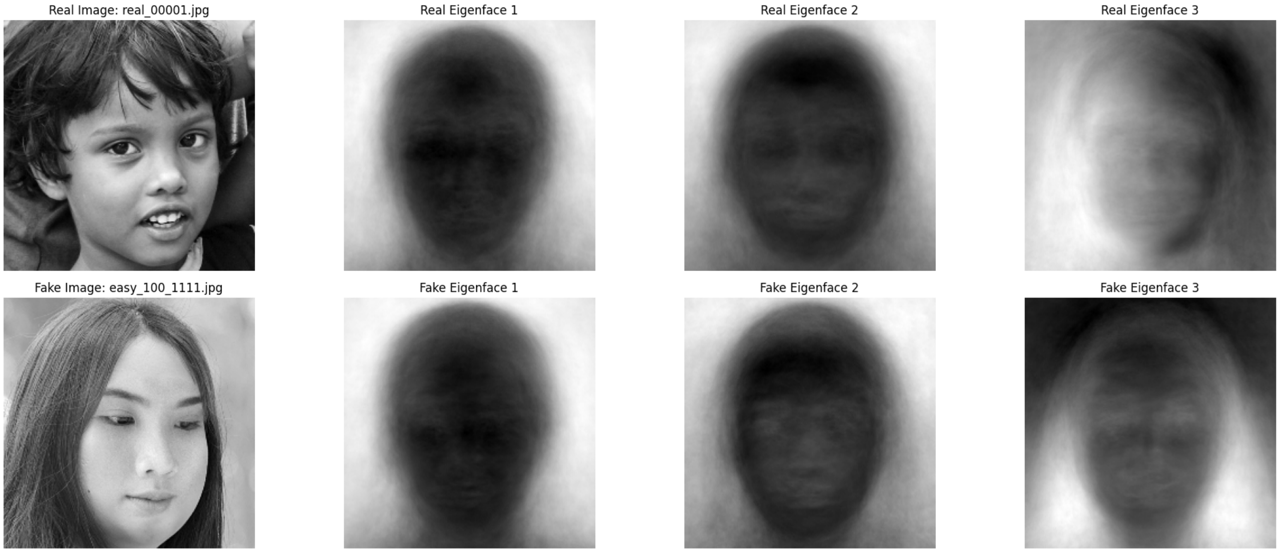

Figure 1.4: Eigen Faces generated by PCA.

1.5 Outline of the remaining chapters

The outline of the remaining chapters is as follows. Chapter 2 provides the recent advancement in the domain of deepfakes. Chapter 3 provides the overall results. Finally, chapter 4 concludes the overall works demonstrated in this study.

Chapter 2

Related Works

Research has progressively utilized AI, expert systems, and neural networks (NNs) to better the detection of counterfeit faces [13]. In research, the generalizability of counterfeit face recognition techniques is evaluated by developing the Fake Face in the Wild (FFW) dataset, comprising over 53,000 images, and examining both local binary patterns with Support Vector Machines (SVMs) and many CNN algorithms, including AlexNet and VGG19 [14]. In another study, Researchers used an NN classifier for identifying counterfeit human faces by utilizing ensemble techniques and image content, achieving AUROC scores of 94% for GAN-generated images and 74.9% for human-created images [15]. A FakeFaceDetect study shows an NN-based system for identifying fake photos generated by GANs, although it fails to reveal the accuracy percentage of its detection efficacy [16]. Advancements in digital manipulation have produced highly realistic counterfeit faces, prompting researchers to utilize DL methodologies, one study commenced with an artificial NN that attained 0.57 accuracy, later improving to 0.77 through the ResNet18 approach for superior classification [17].

Chapter 3

Methodology & Deliverables

3.0.1 Workflow

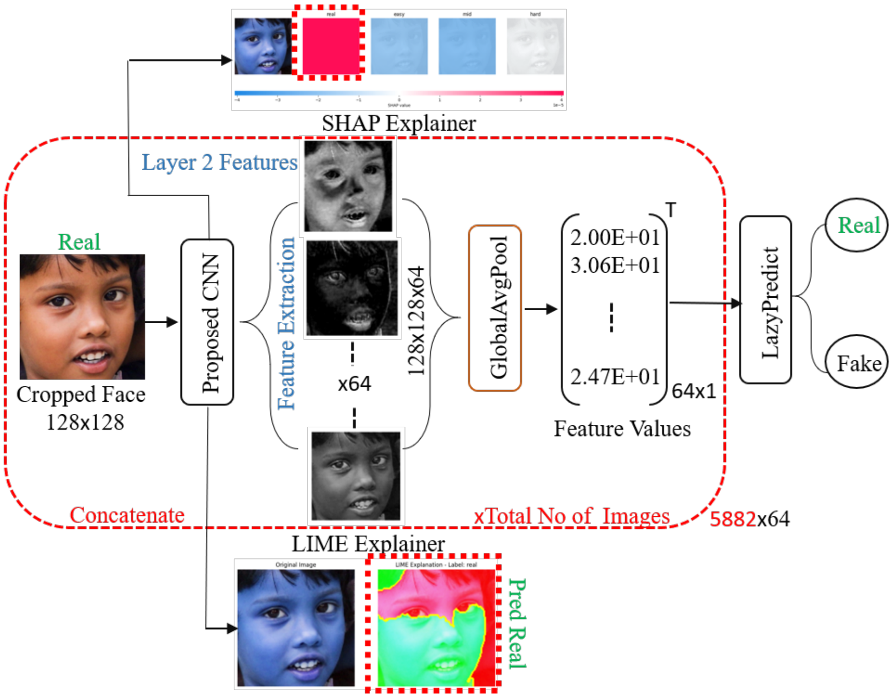

Figure 3.1: An end-to-end CNN-based workflow for distinguishing real from fake images utilizing SHAP and LIME for interpretability, wherein feature extraction, global average pooling, and lazy prediction ascertain authenticity with visual elucidations of critical characteristics.

The original dataset [12] consists of 1081 real images and three categories of a total of 960 fake images—easy, medium, and hard to detect—intended to augment the comprehension of the deep CNN model. Dataset augmentations seek to enhance the model's resilience and generalization by presenting it with a broader spectrum of facial changes and circumstances [18]. To enhance the dataset, we implemented four augmentation techniques: face shifting, width

Algorithm 1 Proposed CNN Model

```
initialize model as Sequential
for filters in {32, 64, 128}:
    add Conv2D layer with filters filters, (3, 3), relu
    add MaxPooling2D layer with (2, 2)
end for
add Flatten layer
add Dense layer with 128 units, relu
add Dense layer with 4 units, softmax
compile the model using 'adam', 'accuracy' and 'categorical_crossentropy'
fit model for 100 epochs and 32 batch size
evaluate model on test data
save model
```

shifting, height shifting, and rotation [6]. A uniform manifold approximation and projection (UMAP) [19] plot is generated to emphasize the distinction between real and fake images at different levels of visibility [see Figure 3.2].

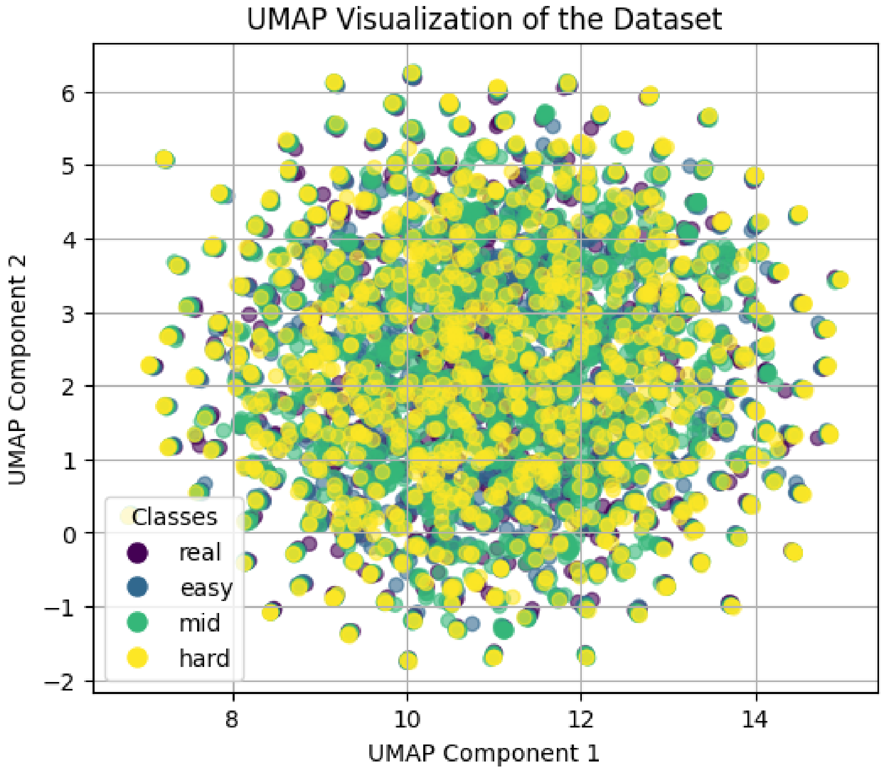

Figure 3.2: The two-dimensional UMAP plot depicting the spread of augmented images across feature space.

Whereas t-SNE moves toward the dimensionality reduction technique called a stochastic neighbor-embedding scheme, which is used for many cases in the visualization of highly dimensional data in two or three dimensions. Figure 3.3 captures the dataset into two components, bringing out the separation according to four classes: real, easy, mid, and hard. Each class is also shown in different colors to assist in finding their overlaps and possible clusters.

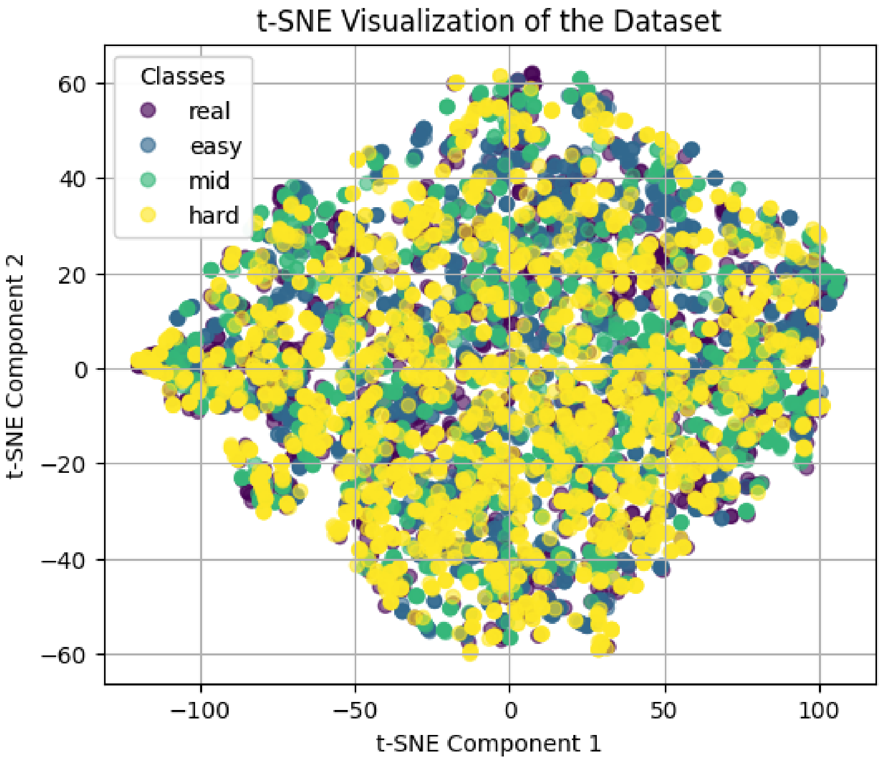

Figure 3.3: The two-dimensional t-SNE plot depicting the spread of augmented images across feature space.

3.0.2 SHAP Explainer

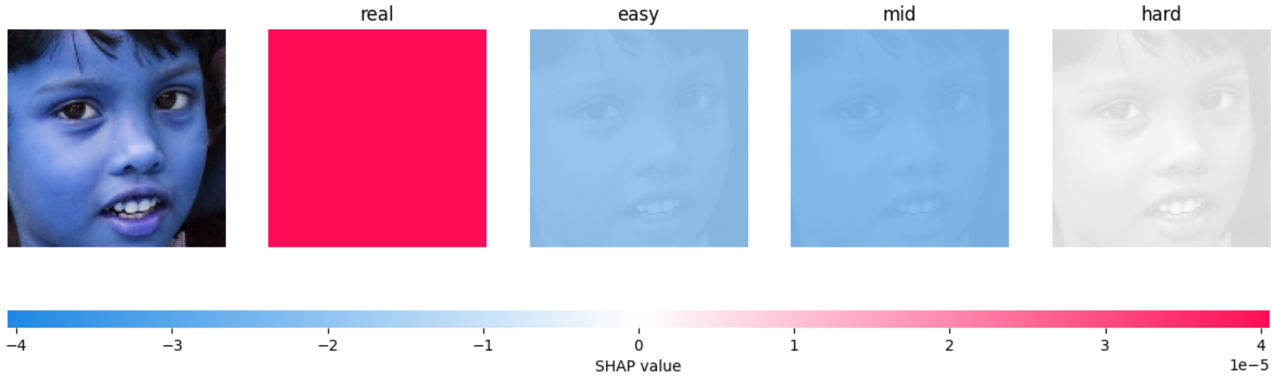

Figure 3.4: SHAP-based accessibility visualization displaying differing confidence levels in predicting facial reliability, with color gradients reflecting feature relevance.

SHAP is an effective tool that expounds AI model predictions by keeping the role of each feature to the entire output predictions [10]. The demonstration provides SHAP explanations for deciding face authenticity, exhibiting a variety of model confidence for each prediction (real, easy, mid, hard). The SHAP values represent feature contributions for identifying each of the classes, with blue denoting negative impact and red indicating positive impact on the model's choice [see Figure 3.4].

3.0.3 LIME Explainer

LIME elucidates the extent to which individual features influence a particular image by the model's predictions [9]. LIME interprets the model's thinking by perturbing segments of the image and analyzing the resultant changes in predictions, therefore identifying regions that significantly affect the model's [see Figure 3.5].

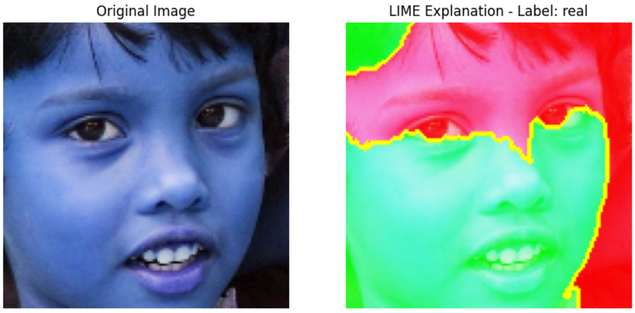

Figure 3.5: LIME-based accessibility visualization depicting facial regions that impacted the model's “real" prediction, with green color-coded areas indicating feature significance for predicting “real" image.

3.0.4 Classification

The LazyPredict method streamlines benchmarking by automatically training and assessing various classifiers, facilitating rapid assessment without considerable optimization [16]. In the workflow [Figure 3.1], following feature extraction via a CNN model and global average pooling, the resultant feature vector is input into LazyPredict, which employs classifiers such as ExtraTrees, RandomForest, LabelPropagation, LabelSpreading, and XGBoost to identify the optimal model for differentiating between real and fake images.

According to the displayed metrics, LazyPredict designates ExtraTreesClassifier as the foremost performer, achieving an accuracy and F1 score of 0.91. This method, using LazyPredict, optimizes model selection, conserving time and resources while improving interpretability via

Table 3.1: Model Performance Metrics

Model	Accuracy	Balanced Accuracy	F1 Score
ExtraTreesClassifier	0.91	0.89	0.91
RandomForestClassifier	0.90	0.88	0.90
LabelPropagation	0.90	0.88	0.90
LabelSpreading	0.90	0.88	0.90
XGBClassifier	0.89	0.87	0.89

SHAP and LIME explainers during the feature extraction phase.

Chapter 4

Conclusion

This work represents the potency of amalgamating feature maps with explainable AI algorithms, like SHAP and LIME, to enhance the precision and clarity of fake face detection methodologies. This approach identifies critical visual patterns by examining feature maps, elucidating their impact on both accurate and inaccurate classifications, and thereby enhancing the understanding of the model's decision-making process. The application of SHAP and LIME enhances the model's robustness and reliability, facilitating the identification of incorrect classification causes and reinforcing the need for further model improvement. This work exposes the relevance of explainable AI in robust detection algorithms by offering vital perceptions for the development of artificial media identification.

Bibliography

[1] Bahar Uddin Mahmud and Afsana Sharmin. Deep insights of deepfake technology : A review, 2023.

[2] Rosa Gil, Jordi Virgili-Gomà, Juan-Miguel López-Gil, and Roberto García. Deepfakes: evolution and trends. *Soft Computing*, 27(16):11295–11318, August 2023.

[3] Rimsha Rafique, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri. Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports*, 13(1):7422, May 2023.

[4] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. pages 4103–4112, 06 2022.

[5] Zhengzhe Liu, Xiaojuan Qi, and Philip Torr. Global texture enhancement for fake face detection in the wild, 2020.

[6] Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing*, 137:104771, 2023.

[7] Fatima Khalid, Ali Javed, Qurat ain, Hafsa Ilyas, and Aun Irtaza. Dfgnn: An interpretable and generalized graph neural network for deepfakes detection. *SSRN Electronic Journal*, 01 2022.

[8] *https://github.com/shankarpandala/lazypredict/tree/master*, accessed 15th Oct, 2024.

[9] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[11] Rami Mubarak, Tariq Alsbou'i, Omar Alshaikh, Isa Inuwa-Dute, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, PP:1–1, 01 2023.

[12] https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection, accessed 20th Sep, 2024.

[13] Sabreen Qwaider, Samy Abu-Naser, and Ihab Zaqout. Artificial neural network prediction of the academic warning of students in the faculty of engineering and information technology in al-azhar university-gaza. pages 16–22, 08 2020.

[14] Tareq Obaid. Factors driving e-learning adoption in palestine: An integration of technology acceptance model and is success model. *SSRN Electronic Journal*, 01 2020.

[15] Noha Eldien, Raghda Ali, and Farid Moussa. Real and fake face detection: A comprehensive evaluation of machine learning and deep learning techniques for improved performance. pages 315–320, 07 2023.

[16] Inés Galván, Jos Valls, Miguel García, and Pedro Isasi. A lazy learning approach for building classification models. *International Journal of Intelligent Systems*, 26:773–786, 08 2011.

[17] Wulan Sapitri, Yesi Novaria Kunang, Ilman Zuhri Yadi, and Mahmud Mahmud. The Impact of Data Augmentation Techniques on the Recognition of Script Images in Deep Learning Models. *Jurnal Online Informatika*, 8(2):169–176, December 2023.

[18] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 07 2019.

[19] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.