Basic Machine Learning approaches for Face Identification on LFW dataset

Gaurav Manish[1] **Hitesh Singh Parihar**[2]
Jaiswal Aditya Ranjit[3] **Ashutosh Kumar**[4]
Vibhor Saxena [6]

[1]**Indian Institute Of Technology,Jodhpur**
{ b22cs079 , b22ee0089 , b22cs025 ,
b22cs015,c23cs1005}@iitj.ac.in

Abstract

In the landscape of computer vision, face identification stands as a pivotal challenge with far-reaching implications across industries. The ability to automatically identify and authenticate individuals from visual data is not only a technological feat but also a cornerstone for numerous applications, from security systems to personalized user experiences.

This project delves deep into the intricacies of face identification, focusing on the application of machine learning algorithms to the Labeled Faces in the Wild (LFW) dataset. Our objective is to dissect the effectiveness of various feature extraction techniques and classification algorithms in accurately recognizing faces from diverse and complex datasets.

At the heart of our investigation lies the fundamental problem of face identification: discerning unique facial characteristics amidst varying lighting conditions, facial expressions, and poses. By tackling this challenge, we aim to unravel the underlying mechanisms that govern the identification process and pave the way for more robust and reliable face identification systems.

Throughout our exploration, we have made several significant findings. Firstly, we discovered that feature extraction techniques such as Convolutional Neural Networks (CNN), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) play a crucial role in capturing distinctive facial features. Moreover, integrating Linear Discriminant Analysis (LDA) with these techniques enhances the discriminative power of the extracted features, leading to improved classification performance.

Our project is structured around these key findings, with each component meticulously designed to explore different facets of the face identification problem. The report is organized into sections corresponding to each stage of our investigation, including data preprocessing, feature extraction, dimensionality reduction, and classification. Within each section, we present our methodologies, experimental results, and critical insights garnered from our analysis.

By elucidating the intricacies of face identification algorithms and their performance under various conditions, this report aims to provide a comprehensive understanding of the state-of-the-art techniques in the field. Our findings not only shed light on the challenges inherent in face identification but also offer valuable insights for researchers and practitioners striving to develop more efficient and reliable face identification systems.

Keywords: LFW , CNN , LBP , HOG

Contents

1 Introduction

Face identification is a fundamental task in computer vision with applications ranging from security systems to social media tagging. In this project, we explore various machine learning algorithms for face identification using the Labeled Faces in the Wild (LFW) dataset. Leveraging the Kaggle API, we access the dataset and preprocess it for feature extraction.

We implement three feature extraction techniques: Convolutional Neural Networks (CNN), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG). Additionally, we integrate Linear Discriminant Analysis (LDA) with these techniques to enhance discriminative power.

Our project consists of six Jupyter Notebooks, each focusing on different combinations of feature extraction and dimensionality reduction tech-

niques. Specifically, we have notebooks for CNN, LBP, and HOG individually, as well as combinations such as CNN with LDA, LBP with LDA, and HOG with LDA.

To evaluate the performance of our models, we employ a variety of classifiers including k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN) using both Scikit-learn and TensorFlow Keras, Random Forest, Support Vector Machines (SVM), Logistic Regression, and Gaussian Naive Bayes. The report is structured to compare the results obtained using different algorithms as well as provide reasons for the obtained results.

Through extensive experimentation and analysis, we compare the accuracy, computational efficiency, and robustness of these algorithms. Our findings provide valuable insights into the effectiveness of different machine learning approaches for face identification tasks.

1.1 Citing paper

Labeled Faces in the Wild (LFW) Dataset:

- "Labeled Faces in the Wild: A Database for Studying Face identification in Unconstrained Environments" by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Available at: http://vis-www.cs.umass.edu/lfw/

Feature Extraction Techniques:

- Convolutional Neural Networks (CNN): Code referenced using the git-hub link provided
- Local Binary Patterns (LBP): Code referenced using the git-hub link provided
- Histogram of Oriented Gradients (HOG): Code referenced using the git-hub link provided

Linear Discriminant Analysis (LDA):

- "Pattern Classification" by Richard O. Duda, Peter E. Hart, and David G. Stork.

Classification Algorithms:

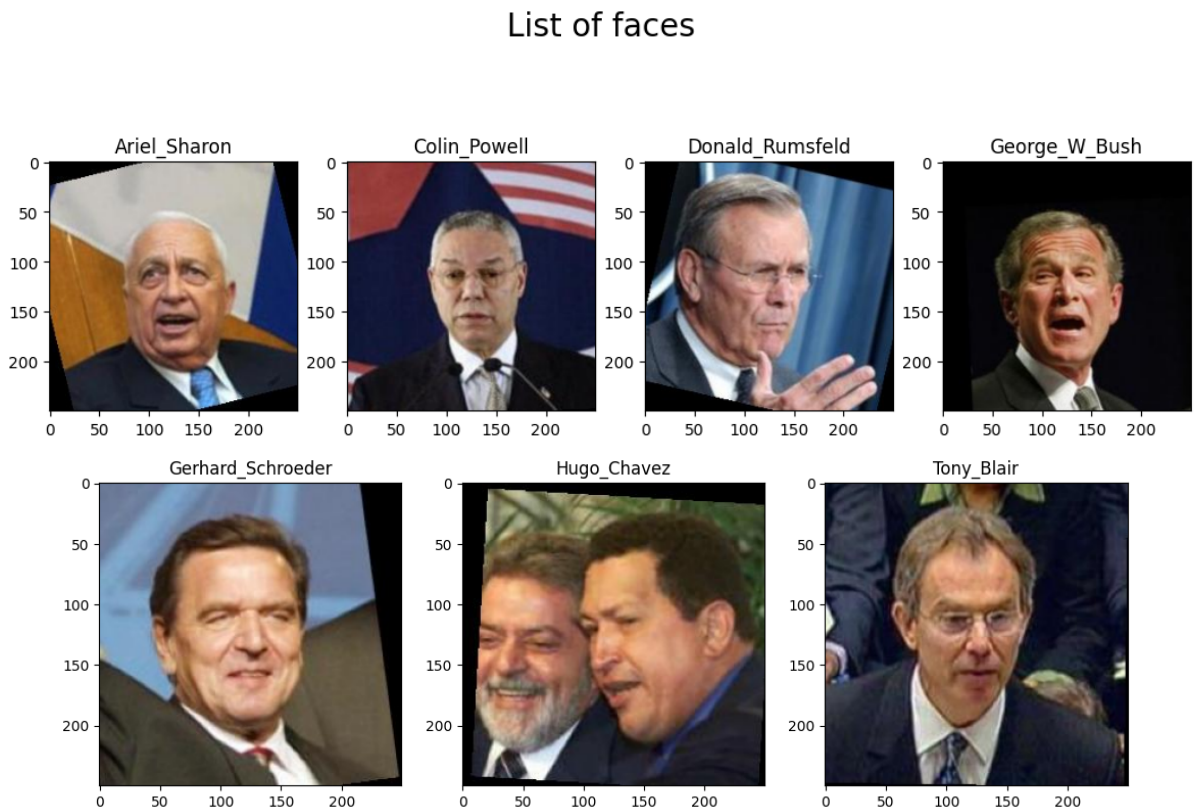
- k-Nearest Neighbors (KNN): "Pattern Classification" by Richard O. Duda, Peter E. Hart, and David G. Stork.
- Artificial Neural Networks (ANN): "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville.

- Random Forest: "Random Forests" by Leo Breiman.
- Support Vector Machines (SVM): "A Tutorial on Support Vector Machines for Pattern Recognition" by Christopher J. C. Burges.
- Logistic Regression: "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
- Gaussian Naive Bayes: "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.

Software Libraries:

- Scikit-learn: "Scikit-learn: Machine Learning in Python" by Fabian Pedregosa et al.
- TensorFlow: "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems" by Martín Abadi et al.

1.2 Figures



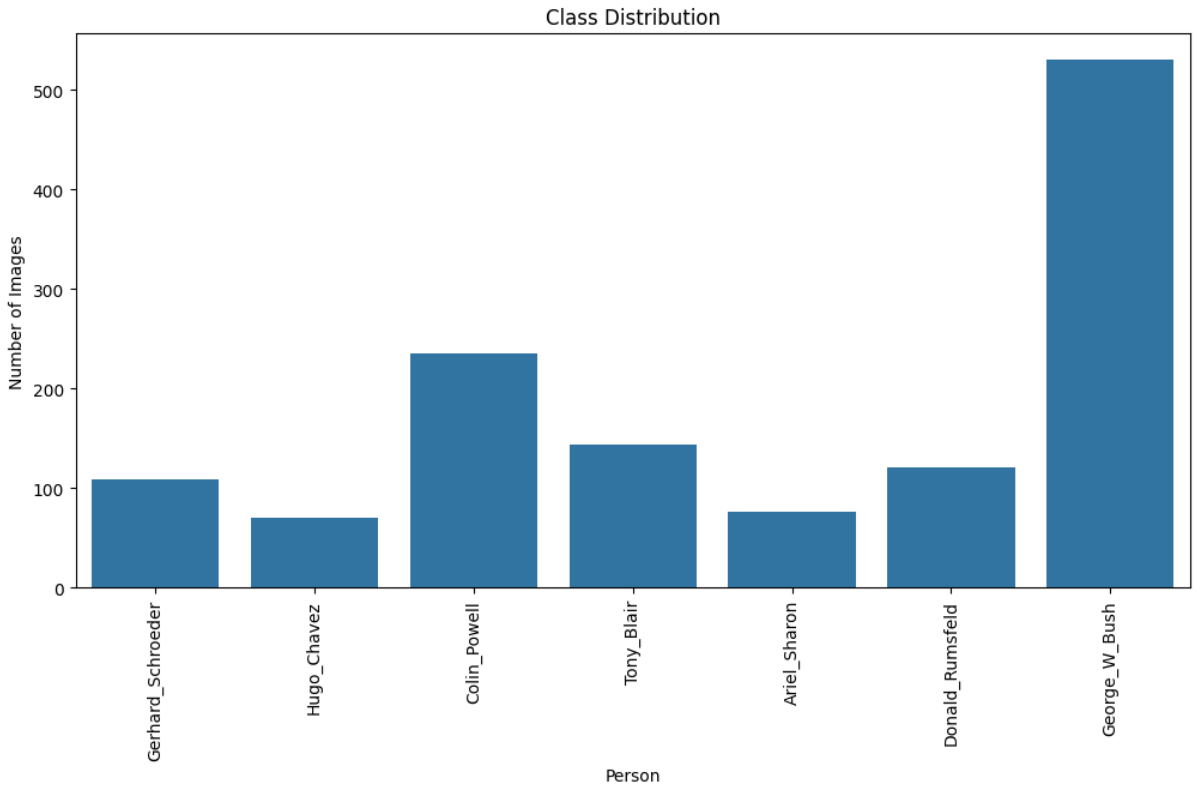


Figure 1: VISUALISING COUNTS

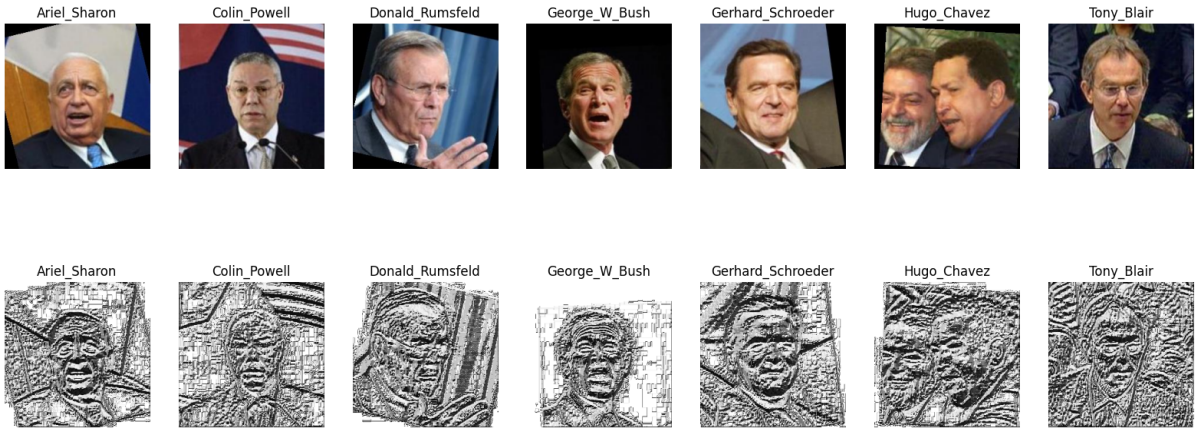


Figure 2: VISUALISING LBP IMAGES

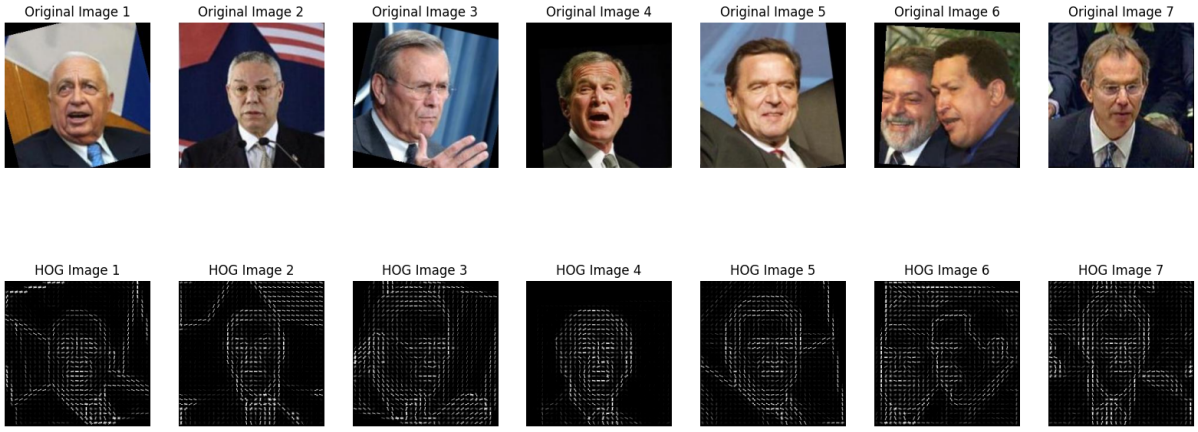


Figure 3: VISUALISING HOG FEATURES

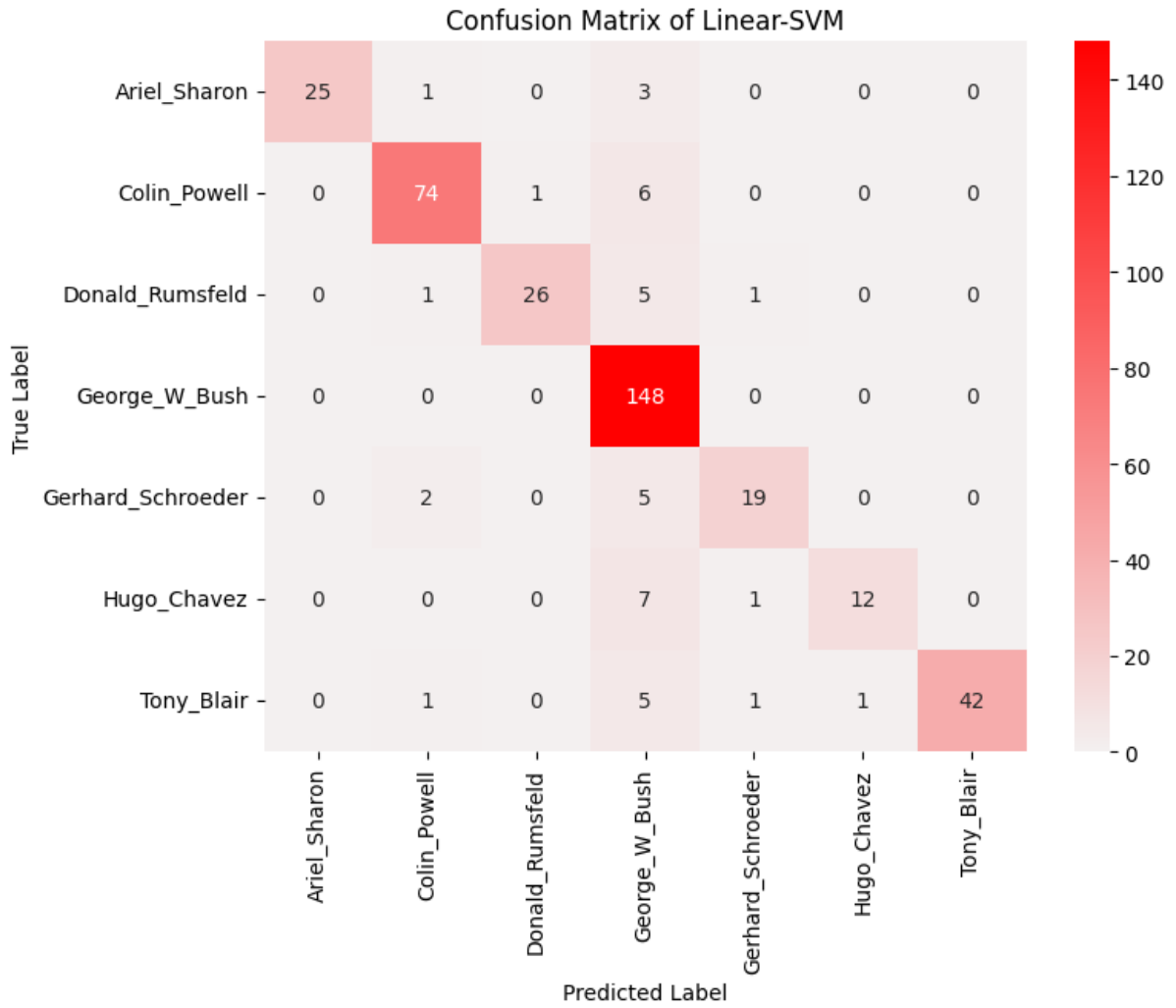


Figure 4: LINEAR SVM ON HOG WITHOUT LDA HAS HIGHEST ACCURACY

2 Approaches Tried

Approach 1: CNN

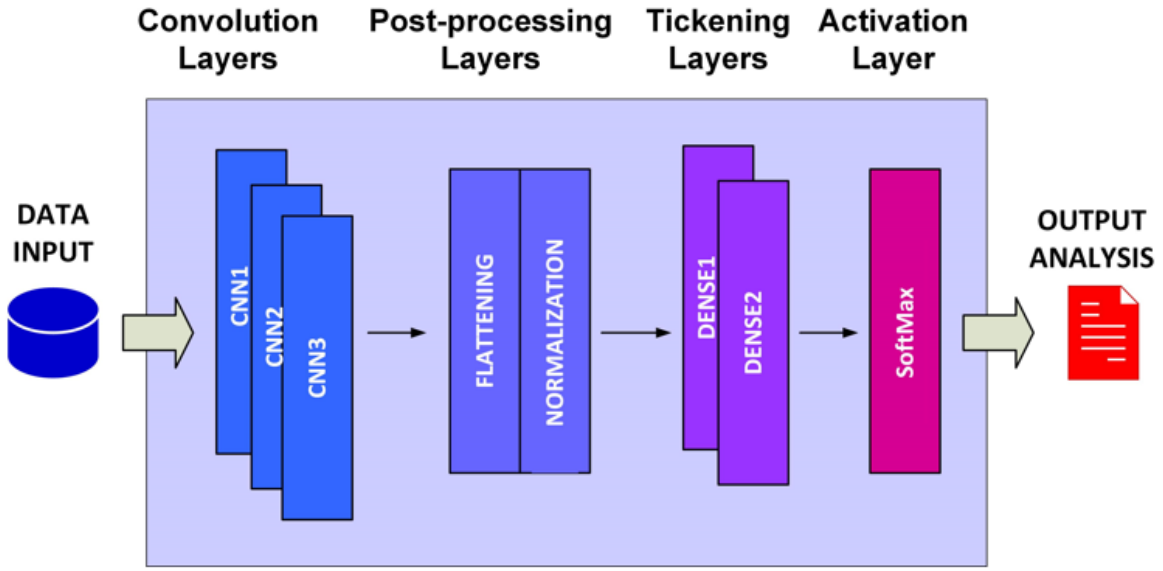


Figure 5: CNN ARCHITECTURE

- We apllied CNN to extract 2048 features for each image and then applied basic ML algorithms such as KNN , ANN , Random Forest , Logistic Regression , Gaussian Naive bias.

Approach 2: LBP

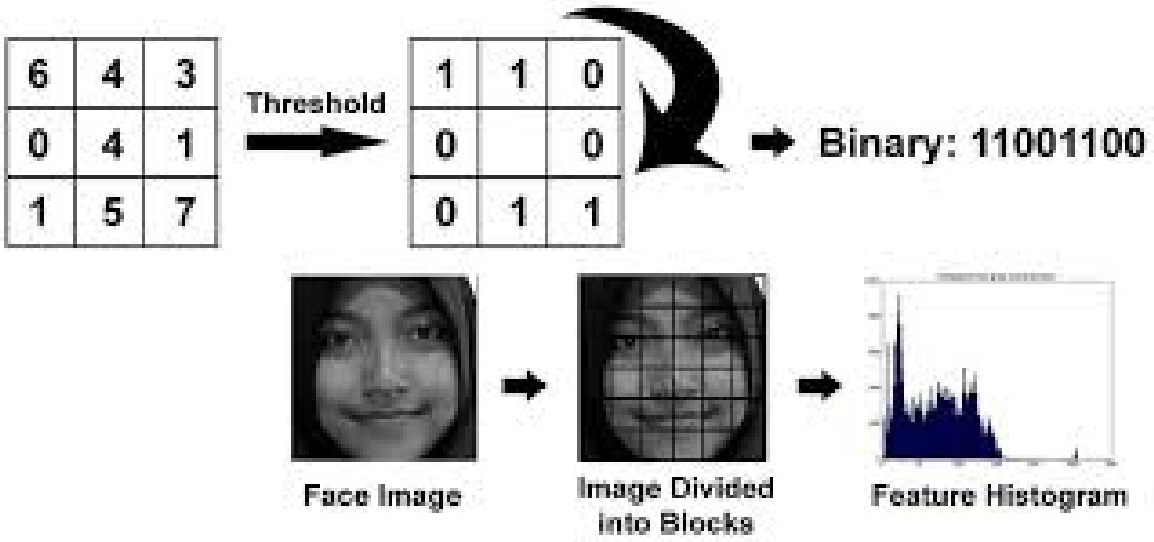


Figure 6: LBP ARCHITECTURE

- We apllied LBP to extract 256 features for each image and then applied basic ML algorithms such as KNN , ANN , Random Forest , Logistic Regression , Gaussian Naive bias.

Approach 3: HoG

- We apllied HoG to extract 70,308 features for each image and then applied basic ML algorithms such as KNN , ANN , Random Forest , Logistic Regression , Gaussian Naive bias.

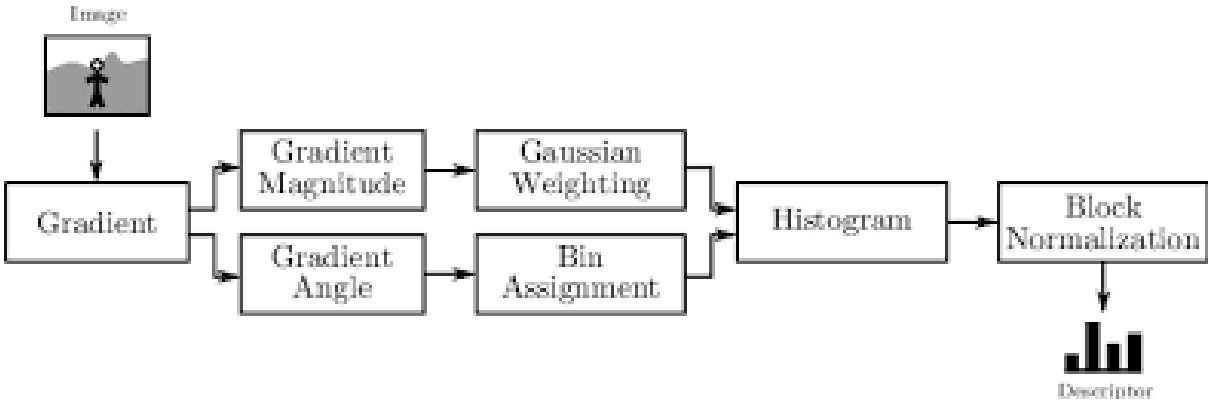


Figure 7: HOG ARCHITECTURE

Approach 4: CNN with LDA

- We apllied CNN to extract 2048 features for each image , then LDA is used to reduce the dataset into a coordinate space maximizing between class scatter and minimizing within class scatter which resulted in reducing the number of dimensions to 6 as there were 7 labels in our data and finally we applied basic ML algorithms such as KNN , ANN ,Random Forest , Logistic Regression , Gaussian Naive bias on the reduced dataset.

Approach 5: LBP with LDA

- We utilised the Local Binary Patterns (LBP) technique to extract 256 features from each image. Subsequently, Linear Discriminant Analysis (LDA) was employed to reduce the dataset by maximising the scatter between classes and minimising the scatter within classes. This reduction resulted in reducing the number of dimensions to 6, considering the presence of 7 labels in our data. Finally, we applied various fundamental machine learning algorithms, including K-Nearest Neighbours (KNN), Artificial Neural Networks (ANN), Random Forest, Logistic Regression, and Gaussian Naive Bayes, on the reduced dataset.

Approach 6: HoG with LDA

- First, we used HoG to get 70,308 features for each image. Next, we used LDA to shrink the dataset into a coordinate space that maximised between class scatter and minimised within class scatter. This cut the number of dimensions to 6 because our data had 7 labels. Finally, we used basic machine learning algorithms like KNN, ANN, Random Forest, Logistic Regression, and Gaussian Naive bias on the smaller dataset..

The Concepts used in our project and some brief details about them has been provided below:

(A) **Convolutional Neural Networks (CNN):** CNNs are deep learning models specifically designed for processing structured grid data, such as images. They consist of multiple layers of convolutional filters followed by pooling layers, allowing them to automatically learn hierarchical representations of features from raw pixel values.

(B) **Local Binary Patterns (LBP):** LBP is a texture descriptor used for texture classification in images. It encodes local texture patterns by comparing each pixel with its neighboring pixels, resulting in a binary pattern representation. LBP is effective in capturing texture variations in images.

(C) **Histogram of Oriented Gradients (HOG):** HOG is a feature descriptor used for object detection and recognition in images. It computes the distribution of gradient orientations in localized regions of an image. HOG is particularly useful for capturing shape and edge information in images.

(D) **Linear Discriminant Analysis (LDA):** LDA is a dimensionality reduction technique used to find the linear combinations of features that best separate different classes in a dataset. It aims to maximize the between-class scatter while minimizing the within-class scatter, leading to a more discriminative feature space.

(E) **k-Nearest Neighbors (KNN):** KNN is a non-parametric classification algorithm that classifies data points based on the majority vote of their nearest neighbors. It makes predictions by identifying the k nearest data points in the feature space and assigning the class label that is most common among them.

(F) **Artificial Neural Networks (ANN):** ANN is a class of machine learning models inspired by the structure and function of biological neural networks. They consist of interconnected nodes organized in layers, with each node performing a simple computation. ANN can learn complex patterns from data and are widely used for classification tasks.

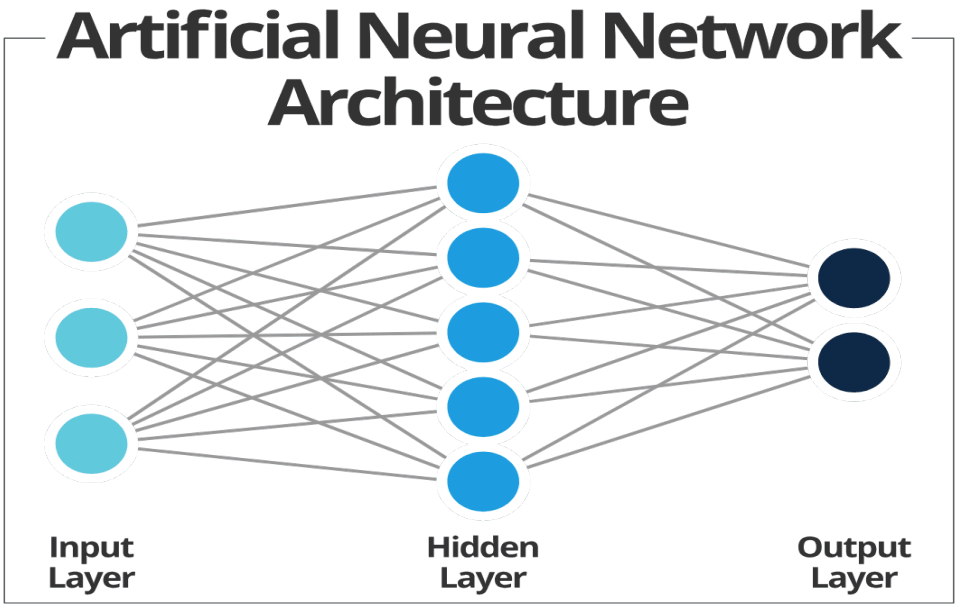


Figure 8: ANN ARCHITECTURE

(G) **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It is known for its robustness and ability to handle high-dimensional data.

(H) **Support Vector Machines (SVM):** SVM is a supervised learning algorithm used for classification and regression tasks. It constructs a hyperplane in a high-dimensional feature space that separates data points into different classes with the maximum margin. SVM is effective in handling both linear and non-linear classification tasks.

(I) **Logistic Regression:** Logistic Regression is a linear classification algorithm used to model the probability of a binary outcome based on one or more predictor variables. It estimates the parameters of a logistic function that maps input features to the probability of the output class.

(J) **Gaussian Naive Bayes:** Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem and the assumption of feature independence. It models the conditional probability of each class given the input features and predicts the class with the highest posterior probability.

3 Experiments and Results

The Labeled Faces in the Wild (LFW) dataset is a widely used benchmark dataset in the field of face identification. It consists of more than 13,000 images of faces collected from the internet, representing over 5,000 different individuals. These images are taken under various conditions, including different lighting, facial expressions, poses, and backgrounds, mimicking real-world scenarios.

In our project we are only loading images of people having atleast 70 images in the dataset to reduce the training time . So in our case the total number of classes are 7.

Table 1: Accuracy results for CNN extracted features

Feature Extraction	**Machine Learning Model**	**Accuracy (%)**
CNN	k-Nearest Neighbors (KNN)	56.57
	Artificial Neural Networks (ANN)	79.81
	Random Forest	51.17
	Linear SVM	76.76
	Polynomial SVM	76
	RBF SVM	78.16
	Logistic Regression	80.28
	Gaussian Naive Bayes	39.67
	TensorFlow Keras ANN	71.83

Table 2: Accuracy results for CNN with LDA extracted features

Feature Extraction	**Machine Learning Model**	**Accuracy (%)**
CNN with LDA	k-Nearest Neighbors (KNN)	82.86
	Artificial Neural Networks (ANN)	80.7
	Random Forest	79.34
	Linear SVM	83.09
	Polynomial SVM	80.28
	RBF SVM	84.27
	Logistic Regression	81.92
	Gaussian Naive Bayes	82.86
	TensorFlow Keras ANN	79.57

Table 3: Accuracy results for LBP extracted feature

Feature Extraction	**Machine Learning Model**	**Accuracy (%)**
LBP	k-Nearest Neighbors (KNN)	37.72
	Artificial Neural Networks (ANN)	23.77
	Random Forest	41.86
	Linear SVM	43.4
	Polynomial SVM	40
	RBF SVM	38.2
	Logistic Regression	45.2
	Gaussian Naive Bayes	35.65
	TensorFlow Keras ANN	30.49

Table 4: Accuracy results for LBP with LDA extracted feature

Feature Extraction	**Machine Learning Model**	**Accuracy (%)**
LBP with LDA	k-Nearest Neighbors (KNN)	44.96
	Artificial Neural Networks (ANN)	39.2
	Random Forest	41.6
	Linear SVM	41
	Polynomial SVM	41.8
	RBF SVM	43.6
	Logistic Regression	79.5
	Gaussian Naive Bayes	79.6
	TensorFlow Keras ANN	42.3

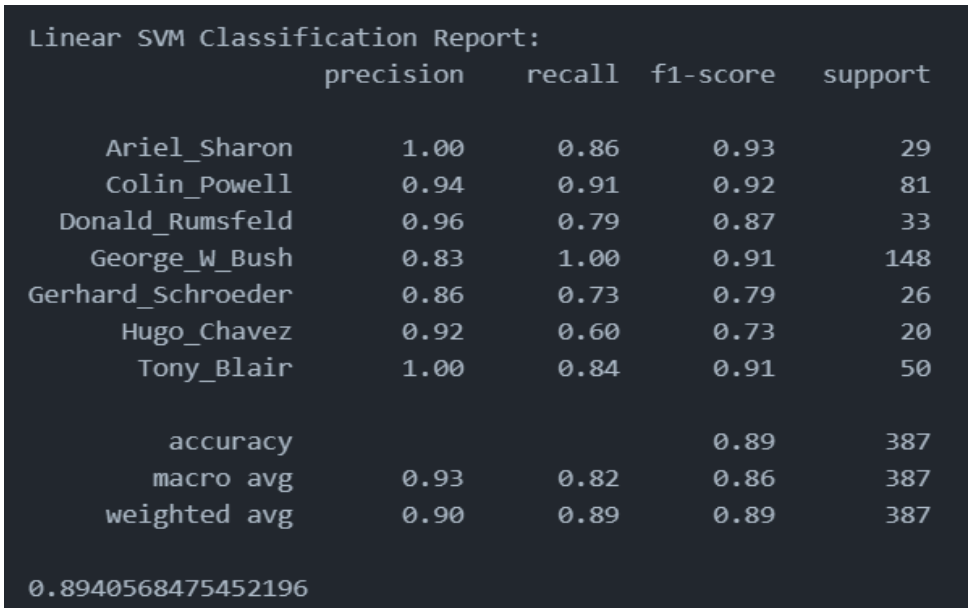
Table 5: Accuracy results for HoG extracted feature

Feature Extraction	**Machine Learning Model**	**Accuracy (%)**
HoG	k-Nearest Neighbors (KNN)	55.81
	Artificial Neural Networks (ANN)	83.2
	Random Forest	63.56
	Linear SVM	**89.40**
	Polynomial SVM	86
	RBF SVM	70
	Logistic Regression	88.3
	Gaussian Naive Bayes	69.5
	TensorFlow Keras ANN	83.2

Table 6: Accuracy results for HoG with LDA extracted feature

Feature Extraction	Machine Learning Model	Accuracy (%)
Hog with LDA	k-Nearest Neighbors (KNN)	79.32
	Artificial Neural Networks (ANN)	82.4
	Random Forest	82.17
	Linear SVM	82.9
	Polynomial SVM	72.35
	RBF SVM	82.68
	Logistic Regression	86.04
	Gaussian Naive Bayes	82.1
	TensorFlow Keras ANN	87

4 Model with best accuracy:



Linear SVM Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	1.00	0.86	0.93	29
Colin_Powell	0.94	0.91	0.92	81
Donald_Rumsfeld	0.96	0.79	0.87	33
George_W_Bush	0.83	1.00	0.91	148
Gerhard_Schroeder	0.86	0.73	0.79	26
Hugo_Chavez	0.92	0.60	0.73	20
Tony_Blair	1.00	0.84	0.91	50
accuracy			0.89	387
macro avg	0.93	0.82	0.86	387
weighted avg	0.90	0.89	0.89	387

0.8940568475452196

Figure 9: best model results

Approach : HOG, Model : Linear SVM without using LDA Results: Our experiments yielded a classification accuracy of **0.89** using the linear SVM classifier with HOG features. This indicates a promising performance of the model in distinguishing between different classes in the dataset. further analysis of other metrics such as precision, recall, and F1-score provided a more comprehensive understanding of the model's performance.

The achieved accuracy of 0.89 demonstrates the potential of the lin-

ear SVM classifier with HoG features for the given task. Fine tuning of hyperparameters is done .

Observations

1. if we look at the dataset it is overloaded with the images of George W Bush.
2. As per the confusion matrix provided above for linear SVM on HOG features it can be seen that highest number of misclassifications are for George W Bush.
3. This is due to the biasing nature dataset as training will be baised toward images of Mr Bush . so any image which is somewhat similar to president George is getting wrongly predicted.

   Linear SVM on HOG features turns out to be best because of following advantages of HOG: Robust feature representation,Dimensionality reduction with effective feature representation,Robustness to noisy data. HOG extracts highest number of features as compared to other two LBP and CNN which will provide more information about data

   After applying LDA on HOG features and using same model , accuracy reduces because:

   i)Loss of information during dimensionality reduction.

   ii)Non-linear relationships in data.

   iii)Sensitive to class imbalance.

5 Summary

The report presents a comprehensive exploration of machine learning algorithms for face identification using the Labeled Faces in the Wild (LFW) dataset. The primary focus is on investigating various feature extraction techniques and classification algorithms to achieve accurate and efficient face identification.

The project begins with data preprocessing, where the LFW dataset is loaded and prepared for feature extraction. Three main feature extraction techniques are explored: Convolutional Neural Networks (CNN), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG). Additionally, Linear Discriminant Analysis (LDA) is integrated with these techniques to enhance feature discriminability and reduce dimensionality.

The report is structured into six distinct sections, corresponding to different combinations of feature extraction and dimensionality reduction methods. Each section includes a detailed methodology, experimental

setup, and analysis of results. Various classification algorithms are applied to evaluate the performance of the extracted features, including k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest, Support Vector Machines (SVM), Logistic Regression, and Gaussian Naive Bayes.

Key findings from the experiments reveal the effectiveness of CNN, LBP, and HOG in capturing discriminative facial features. Furthermore, integrating LDA with these techniques significantly improves classification accuracy and robustness. Comparative analysis of classification algorithms highlights their respective strengths and limitations in the context of face identification tasks.

The report concludes with insights into the most effective combinations of feature extraction and classification algorithms for face identification. It also discusses potential avenues for future research, including exploring advanced deep learning architectures and incorporating additional contextual information for improved performance.

Overall, the report provides valuable insights into the application of machine learning techniques for face identification and lays the groundwork for further advancements in this field.

A Contribution of each member

1. Gaurav Manish(B22CS079): worked on extraction of features , worked on streamlit(backend),workded on the frontend of the website,contributed in report and implemented models.
2. Hitesh Singh Parihar(B22EE089): Helped in applying various models , contributed in preparing report,contributed in video and worked on data preprocessing.
3. Jaiswal Aditya Ranjit(B22CS025):worked on extracting of LBP features , implemented models, exploratory data analysis and contributed to video presentation
4. Ashutosh Kumar(B22CS015): implemented models, Prepared website for the project, worked for the presentation and worked on the deployment of website
5. Vibhor Saxena (C23CS1005): loading of the dataset and implemented few models.