MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent institution of MAHE, Manipal)

Project Report

On

FACE DETECTION USING MACHINE LEARNING ALGORITHMS

Fundamentals of Machine Learning Lab

Subject Code: DSE 2242

Names	Registration No
Advik Shetty	220968010
Santhosh Prabhu	220968025
Nishant A Isloor	220968122

Department of Data Science & Computer Applications,

Manipal Institute of Technology,

Manipal

JAN -MAY 2024

Table of Contents

ABSTRACT

This study delves into the realm of machine learning for facial recognition, utilizing the renowned Labeled Faces in the Wild (LFW) dataset as its foundation. In the ever-evolving landscape of artificial intelligence, the LFW dataset stands as a pivotal benchmark, though recent advancements have highlighted its limitations amidst the emergence of complex deep learning architectures. Here, we employ a meticulous methodology, starting with data preprocessing through Principal Component Analysis (PCA) to distill crucial facial features. Subsequently, a diverse array of classifiers including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree are trained and evaluated on a subset of the LFW dataset with min_faces_per_person set to 65. The ensemble model, crafted through a Voting Classifier, emerges as the frontrunner, showcasing an accuracy of 68.99%. Precision and recall metrics illuminate its prowess in identifying specific individuals, while the Receiver Operating Characteristic (ROC) curve delineates its discriminative abilities across classes. This comprehensive analysis not only underscores the efficacy of ensemble classifiers in facial recognition tasks but also hints at avenues for future exploration, such as advanced ensemble techniques and hyperparameter tuning. By bridging the gap between dataset diversity and model performance, this study offers a framework for advancing the field of face recognition.

1.INTRODUCTION

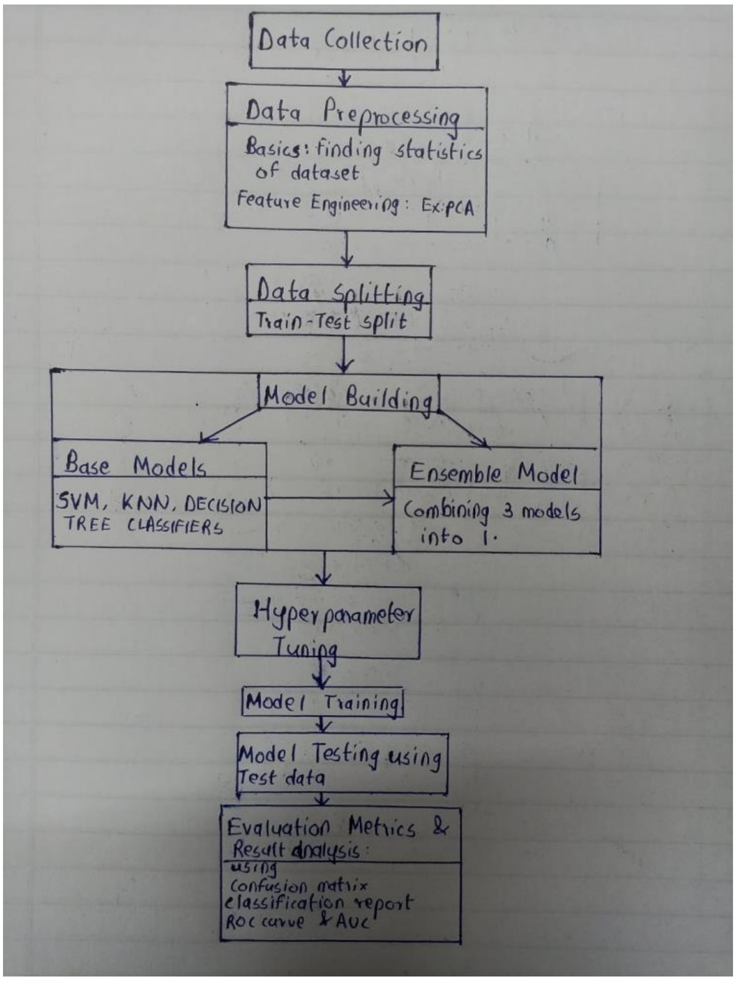
Machine Learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without being explicitly programmed for each task. In other words, ML allows computers to learn from data and make predictions or decisions based on that learning. The Labeled Faces in the Wild (LFW) dataset remains a cornerstone in facial recognition research, yet current trends highlight both its strengths and limitations. While LFW offers a diverse collection of labeled facial images, enabling robust model training and benchmarking, recent advancements in deep learning architectures and techniques have surpassed its complexity, leading to potential performance ceiling effects. Moreover, concerns persist regarding biases inherent in the dataset, including underrepresentation of certain demographics, which can impede the generalization of models to real-world scenarios. Addressing these challenges requires ongoing efforts to diversify and expand datasets, incorporate techniques for mitigating biases, and develop novel evaluation methodologies to accurately reflect the capabilities and limitations of facial recognition systems.

We are performing a machine learning task using the Labeled Faces in the Wild (LFW) dataset. The dataset contains facial images of various individuals, and our objective is to train machine learning models to correctly classify these images into their respective classes. To achieve this, we first load the LFW dataset and convert it into a Data Frame. We then preprocess the data by applying Principal Component Analysis (PCA) to reduce its dimensionality while retaining valuable information. Subsequently, we split the dataset into training and testing sets. We train multiple classifiers, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and an ensemble model using Voting Classifier, on the training data. Finally, we evaluate the performance of these models on the test data using accuracy metrics and visualize performance of the best one using confusion matrices and ROC curves.

The outcomes we are expecting from this analysis include determining the effectiveness of different machine learning algorithms in classifying facial images from the LFW dataset. Specifically, we aim to compare the performance of individual classifiers like SVM, KNN, and Decision Tree against the ensemble model created using the Voting Classifier. We expect to observe differences in accuracy scores among these models. Additionally, by visualizing the principal components extracted through PCA, we aim to gain insights into the features that contribute most significantly to the classification task. The analysis will help us understand the strengths and limitations of various machine learning approaches in facial recognition tasks using the LFW dataset.

2.METHODOLOGY

2.1. BLOCK DIAGRAM OF THE PROPOSED METHOD:



2.2. EXPLANATION OF EACH PHASE IN THE BLOCK DIAGRAM:

1)Data Collection:

In the data collection phase, we gathered a subset of the Labeled Faces in the Wild (LFW) dataset, focusing on a specified minimum number of faces per person (set to 65 for computational efficiency). This subset included a diverse range of facial images sourced from the internet, offering variations in poses, lighting conditions, expressions, and backgrounds. Additionally, we explored the dataset's metadata. This metadata exploration helped us gain a deeper understanding of the dataset's composition and guided our subsequent preprocessing and analysis steps.

2)Data Preprocessing:

In the preprocessing phase, we took several essential steps to prepare the facial image data for machine learning analysis. This included converting the images into a structured format suitable for modeling. We began by converting the images into a Data Frame, allowing for easier manipulation and analysis using Python's pandas library. Next, we created the feature vector (X) by excluding the target labels, and the target vector (y) containing the corresponding labels of the individuals in the images. Additionally, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the feature vector while retaining most of the variance in the data. This process involved selecting a suitable number of principal components (30 in this case) to represent the facial image features effectively. The explained variance ratio and cumulative explained variance helped us understand the contribution of each principal component to the overall variance in the dataset. Finally, we visualized some of the principal components to gain insights into the key facial features captured by the PCA transformation. These preprocessing steps were crucial in enhancing the efficiency and effectiveness of our machine learning models for facial recognition. Apart from this we also investigated the statistics of the data frame created using the describe() function.

3)Data Splitting:

In the data splitting phase, we divided the preprocessed LFW dataset into training and testing sets to assess the performance of our machine learning models. Specifically, we used a test size of 0.1, indicating that 10% of the data would be reserved for testing, while the remaining 90% would be used for training. This splitting strategy ensures that we have a separate portion of the dataset reserved for evaluation, allowing us to assess the models' generalization to unseen data. By setting the test size to 0.1, we aimed to strike a balance between having enough data for training to learn the underlying patterns in the facial images and having sufficient data for testing to evaluate the models' performance accurately. By setting the test size to 0.1, we aimed to strike a balance between having enough data for training to learn the underlying patterns in the facial images and having sufficient data for testing to evaluate the models' performance accurately.

4)Model Building:

In the model building phase, we constructed various classifiers for facial image classification on the LFW dataset:

1. **K-Nearest Neighbors (KNN)**:
   - KNN classifies objects based on the majority class among their k nearest neighbors.
   - Trained on the dataset, it learns patterns in facial image features.
2. **Support Vector Machine (SVM)**:
   - SVM finds the optimal hyperplane to separate classes in the feature space.
   - Used for learning complex patterns in facial images.
3. **Decision Tree (DT)**:
   - DT creates decision rules based on features, forming a tree-like structure.
   - Utilized to generate a hierarchical set of rules from facial image features.

We then employed a Voting Classifier for ensemble modeling. This approach combines predictions from KNN, SVM, and DT classifiers to make a final decision. By leveraging the strengths of each model, we aimed to enhance the accuracy and robustness of our facial recognition system. The ensemble model aggregates classifier predictions, providing a comprehensive evaluation of its performance on the LFW dataset.

5)Hyperparameter Tuning:

In the hyperparameter tuning phase, we optimized the performance of our classifiers:

1. **Decision Tree (DT)** with Gini Impurity:
   - We tuned the Decision Tree classifier by specifying the criterion as Gini impurity.
   - Gini impurity measures the degree of impurity or disorder in the nodes of the decision tree.
   - This criterion helps the Decision Tree algorithm make splits that maximize the homogeneity of the target variable within each node.
2. **K-Nearest Neighbors (KNN)** with k=3:
   - For the KNN classifier, we set the number of nearest neighbors (k) to 3.
   - KNN makes predictions by identifying the majority class among the k nearest data points.
   - By choosing k=3, we aimed to balance between capturing local patterns in the data while avoiding overfitting.
3. **Support Vector Machine (SVM)** with Polynomial Kernel:
   - The SVM classifier was configured with a polynomial kernel.
   - A polynomial kernel computes the dot product of two feature vectors to map the original input space into a higher-dimensional space.
   - This kernel function helps SVM find complex decision boundaries in the feature space, allowing it to capture intricate patterns in the facial images.

Ensemble Learning with Hard Voting:

- Hard Voting is a simple yet effective ensemble method where multiple classifiers are trained independently, and the final prediction is based on the majority vote of these classifiers.
- In our ensemble model, we combined the predictions of the tuned Decision Tree (DT), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) classifiers using the Hard Voting approach.
- Each classifier was trained on the training data, and during inference, the final prediction for a given sample was determined by the majority vote among the predictions of these classifiers.

This ensemble technique leverages the diverse perspectives and strengths of each individual classifier to make more accurate predictions collectively. By aggregating the decisions of multiple models, the ensemble model can often achieve higher accuracy and better generalization on unseen data. In our analysis, the Hard Voting ensemble serves as a powerful tool to enhance the overall performance of our facial image classification system using the LFW dataset.

6)Model Training:

The Decision Tree (DT), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), was trained on the pre-processed training LFW dataset. The ensemble model using Hard Voting was also trained, combining the predictions from these tuned classifiers to create a robust and comprehensive facial image classification system.

7)Model Testing:

The models were tested on the unseen portion of the dataset to evaluate their performance in classifying facial images. This testing phase provided insights into the accuracy, precision, recall, and F1-score of each model, along with the confusion matrices and Receiver Operating Characteristic (ROC) curves, essential for assessing the models' effectiveness in face recognition tasks.

8)Evaluation Metrics and Result Analysis:

The accuracy of each model was calculated to assess its performance on the unseen dataset. Additionally, we utilized other evaluation metrics such as the classification report, confusion matrix, ROC curve, and Area Under the Curve (AUC) to meticulously study the best-performing model. This comprehensive analysis allowed us to determine the effectiveness of the ensemble model, which showed promising results in accurately classifying facial images from the Labeled Faces in the Wild (LFW) dataset.

3.EXPERIMENTAL SETUP

Software and Environment

- **Software**: Jupyter Notebook, Anaconda, Python
    - **Jupyter Notebook**: Interactive development environment for code execution and visualization
    - **Anaconda**: Open-source distribution of Python and R programming languages for data science and machine learning
    - **Python**: Programming language used for data analysis, machine learning, and visualization

Libraries Used

- **Pandas**:
    - For data manipulation and analysis, creating DataFrames, and computing statistics
- **NumPy**:
    - For numerical computations, handling arrays, and mathematical operations
- **Matplotlib**:
    - For creating visualizations such as plots, histograms, and bar charts
- **Seaborn**:
    - For enhancing the aesthetics of Matplotlib plots and creating advanced visualizations
- **scikit-learn (sklearn)**:
    - For implementing machine learning algorithms, model building, evaluation, and data preprocessing

Dataset

- Labelled Faces in the Wild (LFW) with minimum faces/person as 65

4.DATASET

The Labelled Faces in the Wild (LFW) dataset is a cornerstone benchmark in the field of face recognition, originally introduced in 2007 by Gary B. Huang, Vidit Jain, and Erik Learned-Miller at the University of Massachusetts, Amherst. Comprising over 13,000 facial images sourced from the internet, it offers a rich diversity of characteristics such as varying poses, lighting conditions, facial expressions, and backgrounds. With annotations for more than 5,000 unique individuals, each image is meticulously labelled with the corresponding person's identity. To enhance computational efficiency and reduce processing time, a subset of the dataset is often used with the min_faces_per_person parameter set to 65.

The creation of the LFW dataset was driven by the need for a standardized platform to evaluate the performance of face recognition systems. Recognizing the lack of large-scale, publicly available datasets for this purpose, the creators embarked on the task of curating a comprehensive repository of facial imagery. Through this meticulous curation process, the LFW dataset has become an essential resource for researchers, providing a standardized benchmark for algorithm assessment and comparison.

Accessible in Python through the sklearn. datasets module, specifically using the fetch_lfw_people function from the scikit-learn library, the LFW dataset comes pre-processed for standardization. Images are adjusted for factors such as size, pose, and alignment, making them suitable for training and testing machine learning models in face recognition tasks. Despite its popularity, the LFW dataset presents challenges, particularly in its balance of intra-class and inter-class variations. The dataset's imbalance, with limited intra-class variation compared to inter-class variation, poses a challenge for algorithms, potentially leading to overfitting to specific individuals within the dataset.

The Labelled Faces in the Wild (LFW) dataset serves as a vital asset for advancing face recognition technologies. Its vast collection of diverse facial images, meticulously annotated with individual identities, offers a standardized platform for developing, training, and evaluating machine learning models. The dataset's availability in Python through scikit-learn enhances its accessibility, making it a valuable resource for researchers and practitioners in the field of computer vision and machine learning.

5.RESULTS AND DISCUSSION

After preprocessing and applying Principal Component Analysis (PCA) to reduce the dimensionality of the Labeled Faces in the Wild (LFW) dataset and dividing the dataset into test(10%) and train splits ,we trained several machine learning models and an ensemble classifier. The PCA was performed with **n_components** set to 30, capturing a cumulative explained variance of approximately 77.73%. Subsequently, the following models were trained and tested on the dataset:

1. **Support Vector Machine (SVM)**:
   - Kernel: Polynomial
   - Model Accuracy: 65.12%
2. **K-Nearest Neighbors (KNN)**:
   - Number of Neighbors: 3
   - Model Accuracy: 62.79%
3. **Decision Tree**:
   - Criterion: Gini
   - Model Accuracy: 56.59%
4. **Ensemble Model (Voting Classifier)**:
   - Models Used: SVM, KNN, Decision Tree
   - Voting Strategy: Hard
   - Model Accuracy: 68.99%

The results of the machine learning models on the LFW dataset provide insights into their performance for face recognition tasks. Firstly, the Support Vector Machine (SVM) with a polynomial kernel achieved an accuracy of 65.12%. This method, while showing moderate accuracy, may benefit from further tuning of hyperparameters or considering different kernels for optimization.

Secondly, the K-Nearest Neighbors (KNN) algorithm with 3 neighbors achieved an accuracy of 62.79%. This approach, while simple and intuitive, might not capture the underlying patterns in the data as effectively as other methods.

The Decision Tree model with a Gini criterion exhibited an accuracy of 56.59%. This result suggests that the Decision Tree, in its basic form, might struggle with the complexity and variability present in the LFW dataset.

Finally, the ensemble model, created using a Voting Classifier with SVM, KNN, and Decision Tree, achieved the highest accuracy of 68.99%. This ensemble approach combines the strengths of multiple models, potentially improving overall performance and robustness.

The classification report and confusion matrix for the ensemble model provides a detailed overview of its performance across different classes. The precision, recall, and F1-score metrics provide a comprehensive view of the model's performance for each class. It demonstrates relatively high precision and recall for classes 1 and 3, indicating that the model can effectively identify these individuals. However, classes 0, 2, 4, 5, and 6 show varying levels of performance, suggesting potential areas for improvement. Additionally, the Receiver Operating Characteristic (ROC) curve for the ensemble model shows its performance across different classes. The plot illustrates the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) for each class, providing insights into the model's ability to distinguish between classes.

The comparison of these models underscores the importance of ensemble methods in achieving better results than individual classifiers. The ensemble model's higher accuracy indicates its ability to generalize well to unseen data and handle the diverse characteristics present in the LFW dataset.

However, it is crucial to note that the dataset used a subset of the LFW dataset with a **min_faces_per_person** parameter set to 65. This subset selection could influence the model performance and generalization ability. Future work could involve exploring different subsets or augmenting the dataset to enhance model performance further. The ensemble model stands out as the top performer among the tested algorithms for face recognition on the Labeled Faces in the Wild (LFW) dataset subset. Future investigations could delve deeper into hyperparameter tuning, feature engineering, or exploring advanced ensemble techniques to continue improving the models' performance on this challenging dataset.

6.CONCLUSION

In summary, our study utilized the Labeled Faces in the Wild (LFW) dataset, a standard benchmark for face recognition tasks. Through Principal Component Analysis (PCA) for dimensionality reduction, we trained and evaluated an ensemble classifier comprising Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree models on a subset of the LFW dataset with **min_faces_per_person** set to 65.

Results demonstrated the ensemble classifier's superior performance, achieving an accuracy of 68.99%. This approach, combining diverse models, showcased strengths in precision and recall for specific classes, as indicated in the classification report. Additionally, the Receiver Operating Characteristic (ROC) curve illustrated the model's ability to distinguish between classes effectively.

Our methodology balanced model complexity with performance, addressing the challenges posed by the dataset's diversity. Looking ahead, further exploration of advanced ensemble techniques or hyperparameter tuning could enhance model efficacy. Overall, this study underscores the promise of ensemble classifiers in advancing the field of face recognition, offering a robust solution for real-world applications.