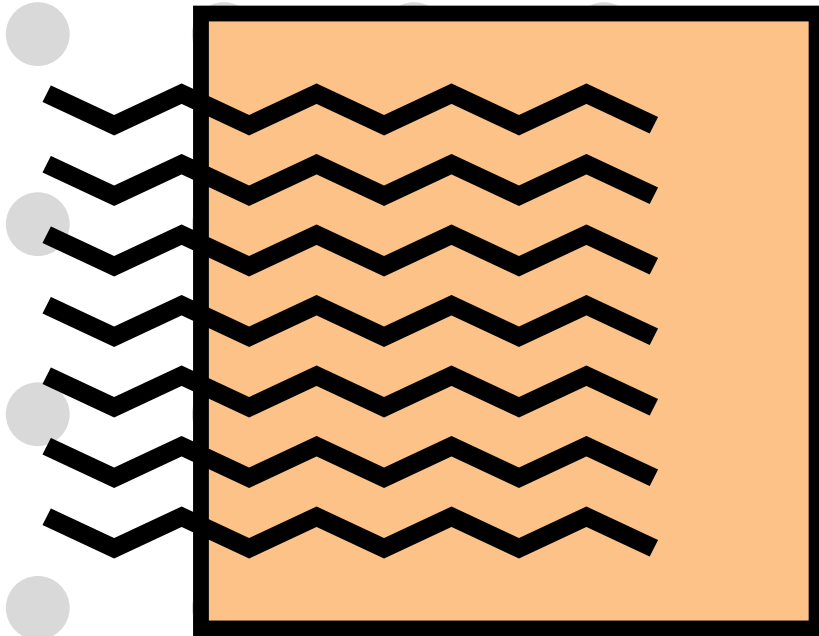


TEAM 5



OCR CASE STUDY

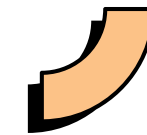


Srihitha Mallepally
Komal Gupta



HELLO & WELCOME

Welcome to our presentation on character classification using machine learning. Our study focused on accurately identifying characters from images and evaluating model performance. Today, we'll share our findings, insights, and recommendations. Let's dive into the details and explore the outcomes of our experiments.



LETS GET STARTED





LIST OF CONTENT PRESENTATION

01

FLOWCHART
FOR ML PIPELINE

02

COMPARISON
OF THE THREE
MODELS

03

THE BEST MODEL
AND
HYPERPARAMETER
TUNING

04

CONFUSION
MATRIX

05

LOW
ACCURACY

06

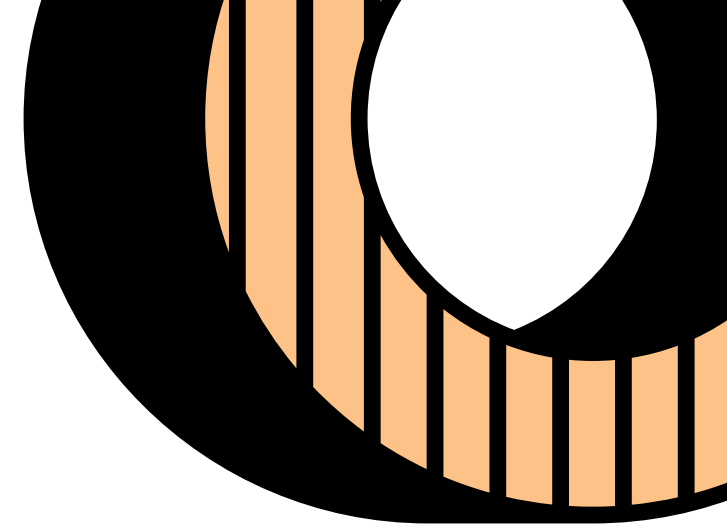
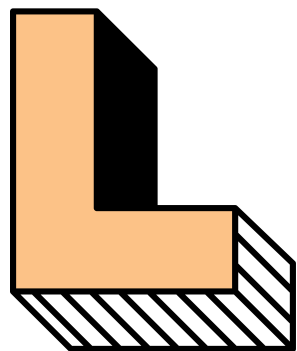
OTHER
ALGORITHMS TO
IMPROVE
ACCURACY:
NEURAL
NETWORKS

07

CHALLENGES
FACED

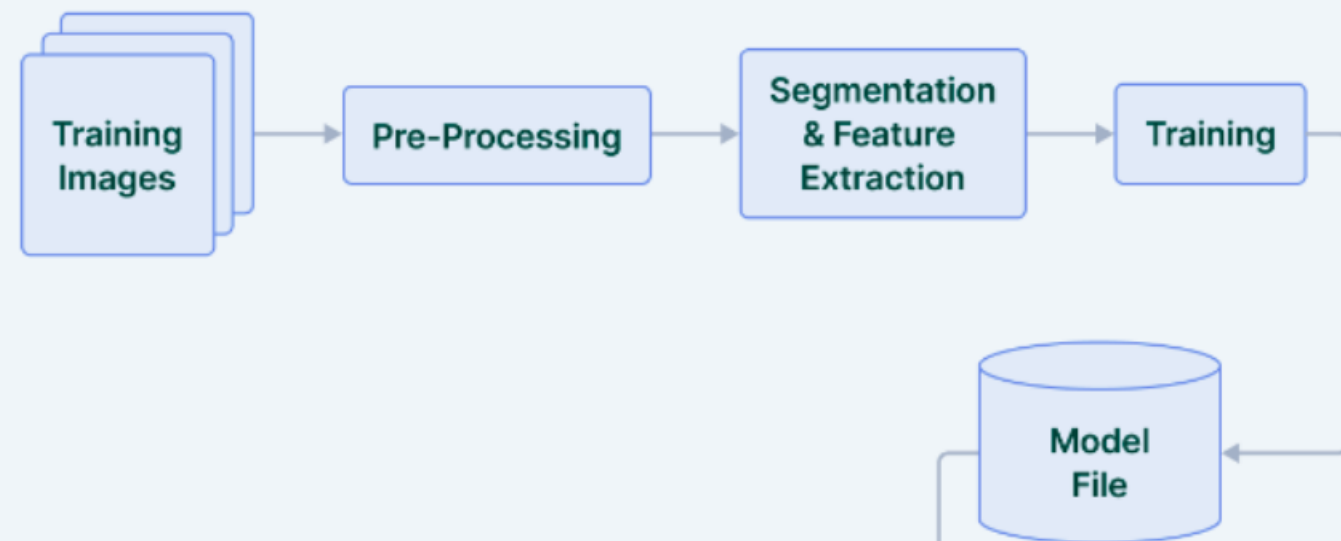
08

CONCLUSIONS

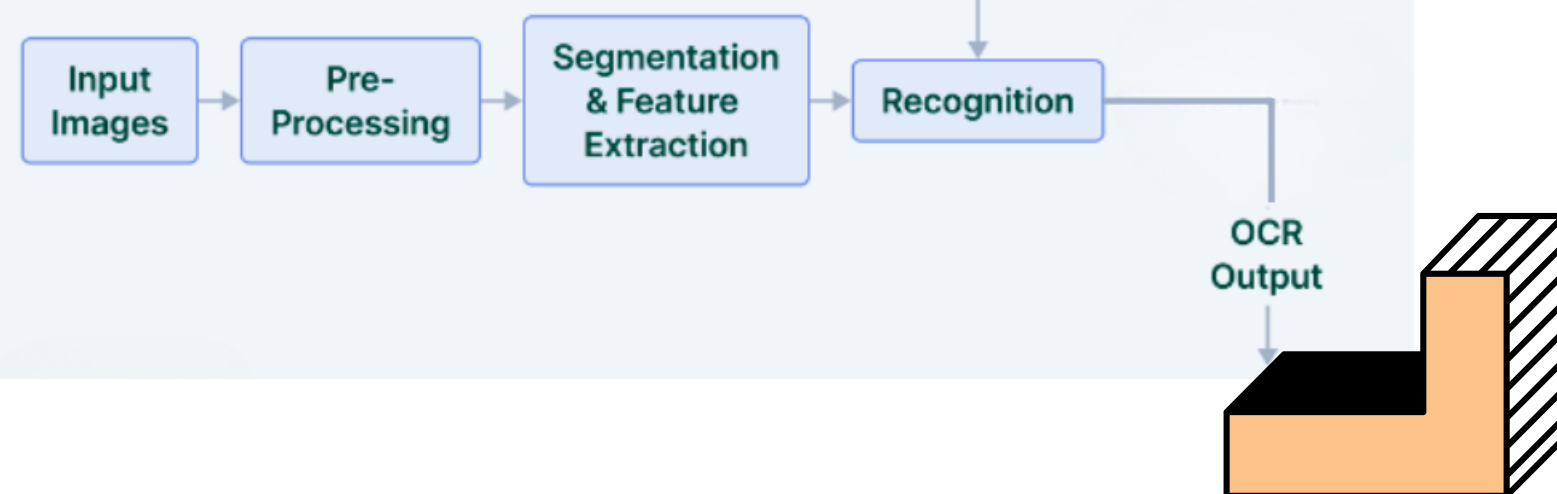


General OCR model

Training Pipeline



Testing Pipeline



FLOW CHART FOR ML PIPELINE

Optical Character Recognition (OCR) is used to process images or scanned documents to produce raw text or other structured output.

STEPS:

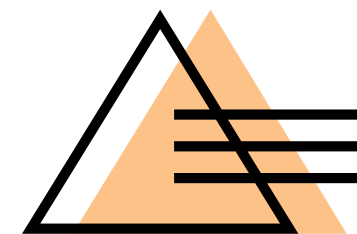
- Image Preprocessing: This step involves preprocessing the input image to enhance its quality and improve OCR accuracy. Common preprocessing techniques include resizing, noise removal, contrast adjustment, and image normalization.
- Feature Extraction: In this step, relevant features are extracted from the preprocessed image. These features could include edges, lines, corners, or other visual patterns that can aid in character recognition.
- Text Recognition: This is the core step where the machine learning model processes the extracted features and recognizes the characters present in the image.
- Output: The final output of the OCR pipeline is the extracted text from the input image. This text can be further used for various purposes such as data analysis, indexing, or information retrieval.



COMPARISON OF THE THREE MODELS

BASED ON ACCURACY SCORE:-

Accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions made. We calculate it by dividing the number of correct predictions by the total number of predictions. The “Random Forest Classifier” yields the best accuracy on the validation set.



Accuracy

Logistic Regression

0.578

SVMs

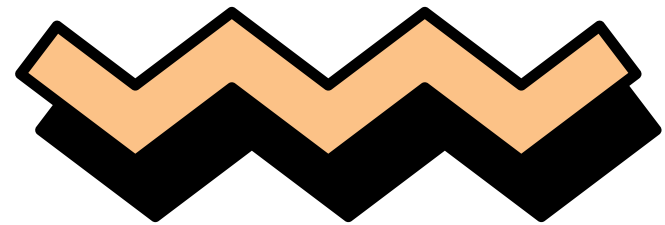
0.736

Random Forest

0.746



THE BEST MODEL AND HYPERPARAMETER TUNING



The Random Forest model performed the best.

There are many hyperparameters for Random Forest. For example, `n_estimators`, `max_depth`, `max_sample_split`, `criterion`.

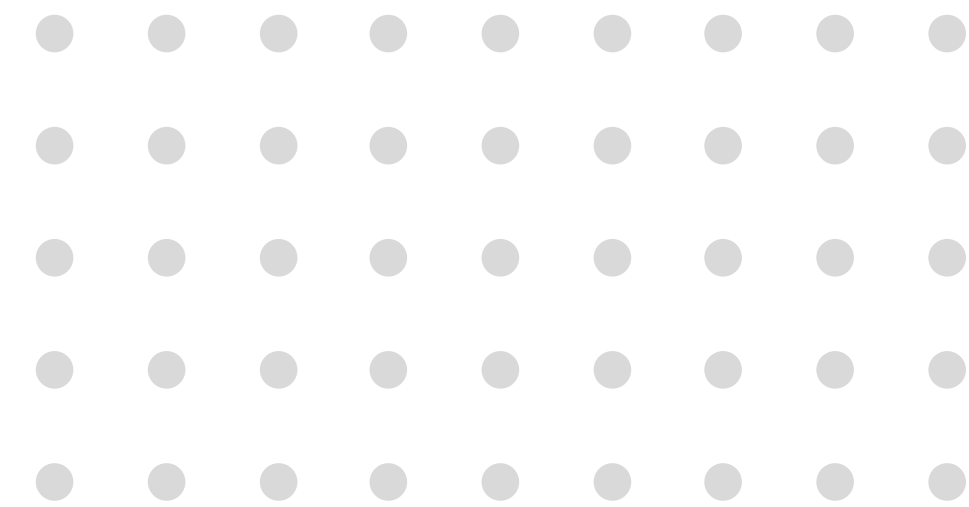
Optimal values should be selected for maximum performance.

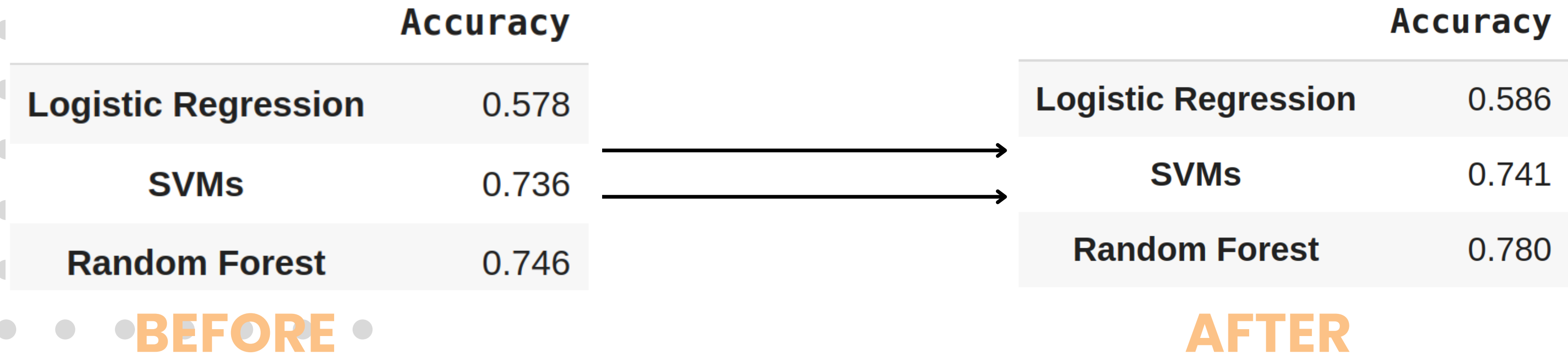
These optimal values can be selected using search methods like grid search or randomised search.

The optimal values found are: `n_estimators=1000`.

We can also modify the train-test splits to increase the accuracy.

```
{'criterion': 'gini', 'min_samples_split': 2, 'n_estimators': 1000}
```



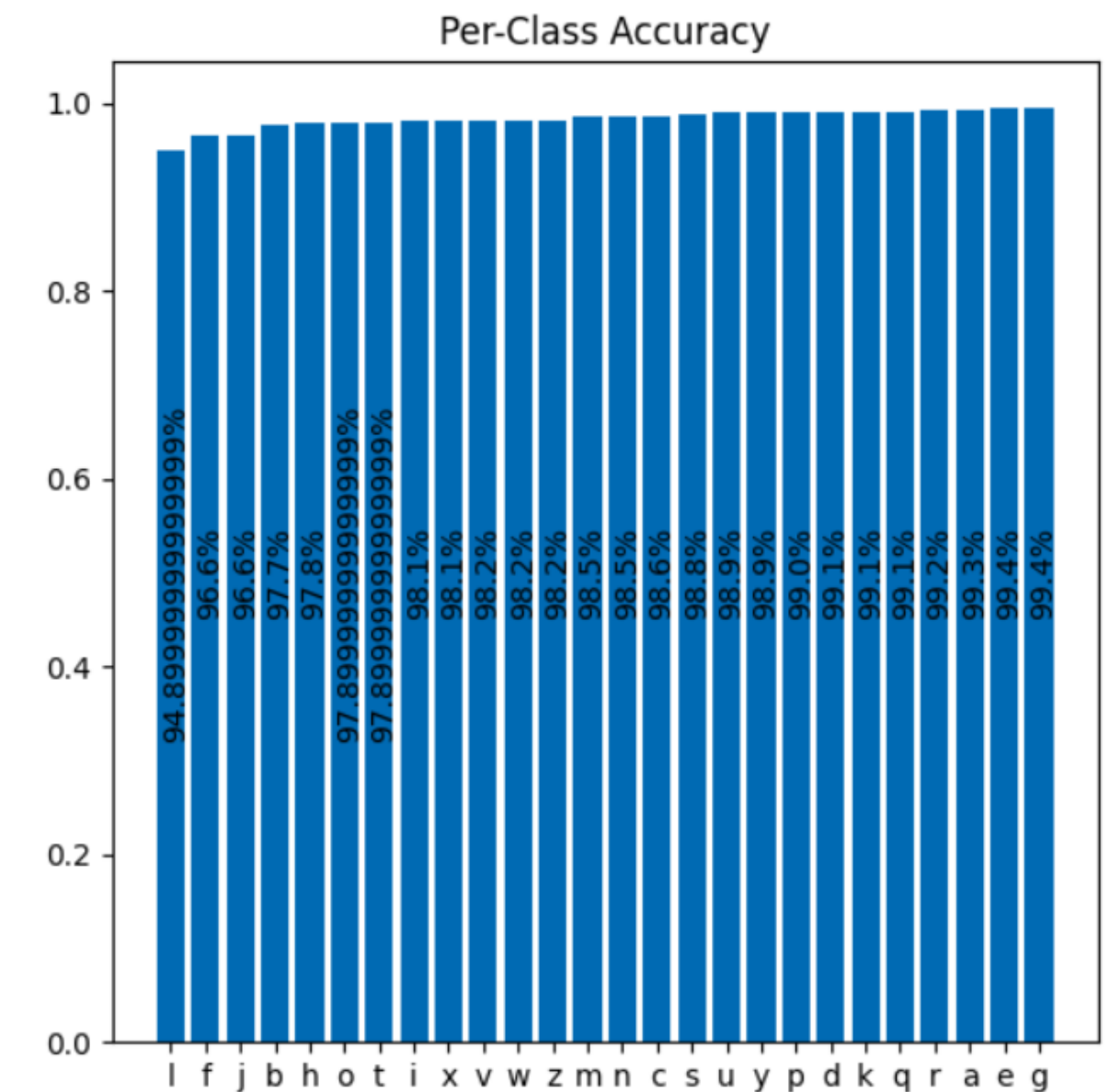
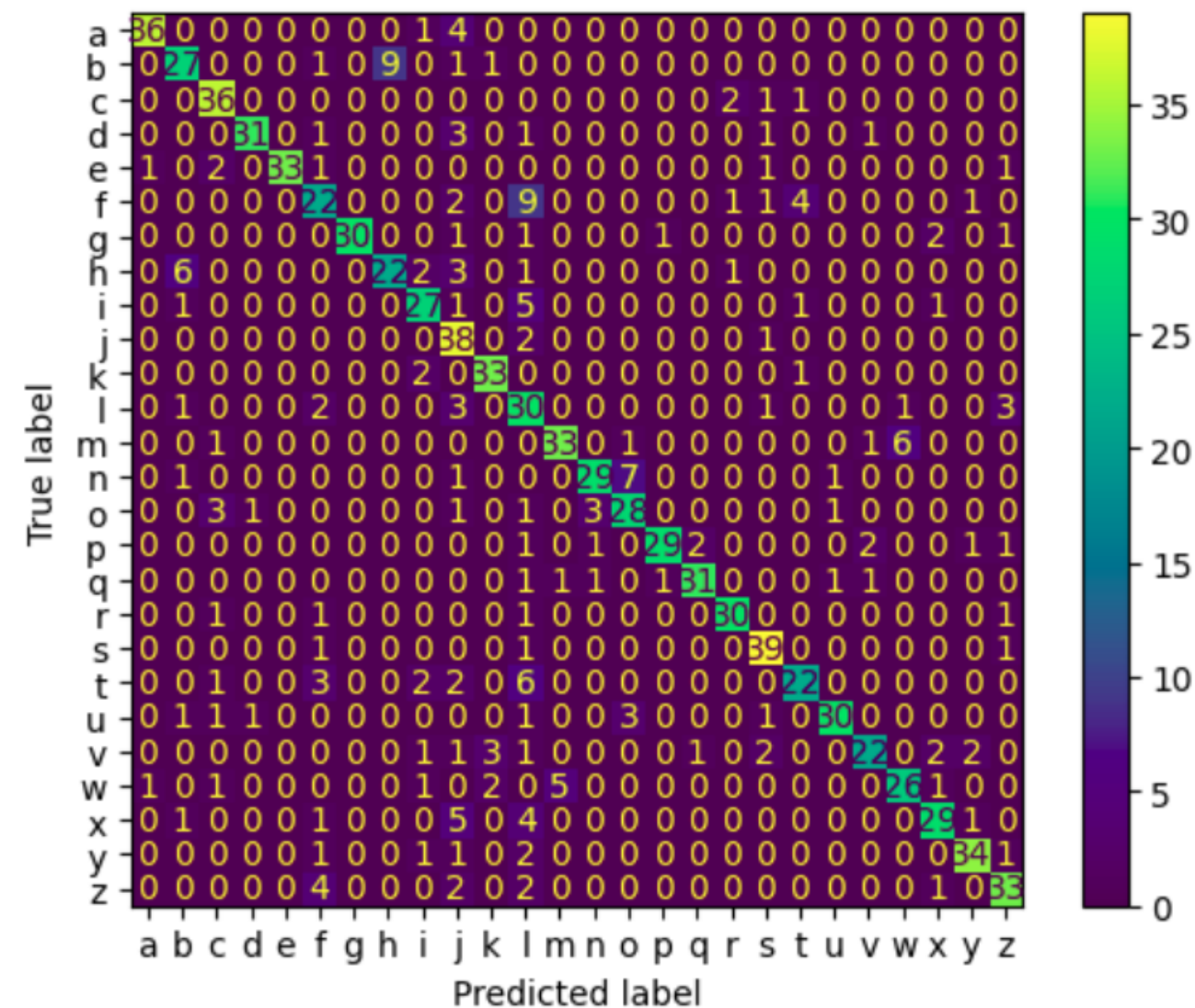


	Precision	Recall	F1 Score
Logistic Regression	0.589437	0.586	0.583837
SVMs	0.776656	0.741	0.747932
Random Forest	0.802417	0.780	0.784797



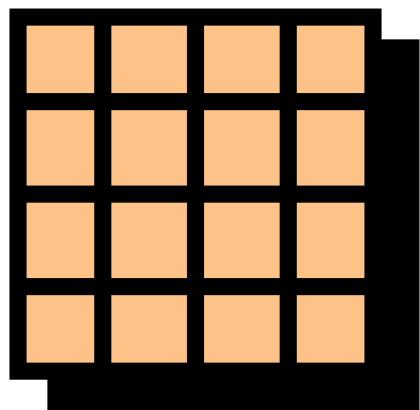
CONFUSION MATRIX

Confusion Matrix using the Random Forest Classifier `RanFor_model(n_estimators=1000)`



$$accuracy = \frac{tp + tn}{total}, \quad tp = CM_{ii} \quad tn = \sum_{j,k \neq i} CM_{jk}$$

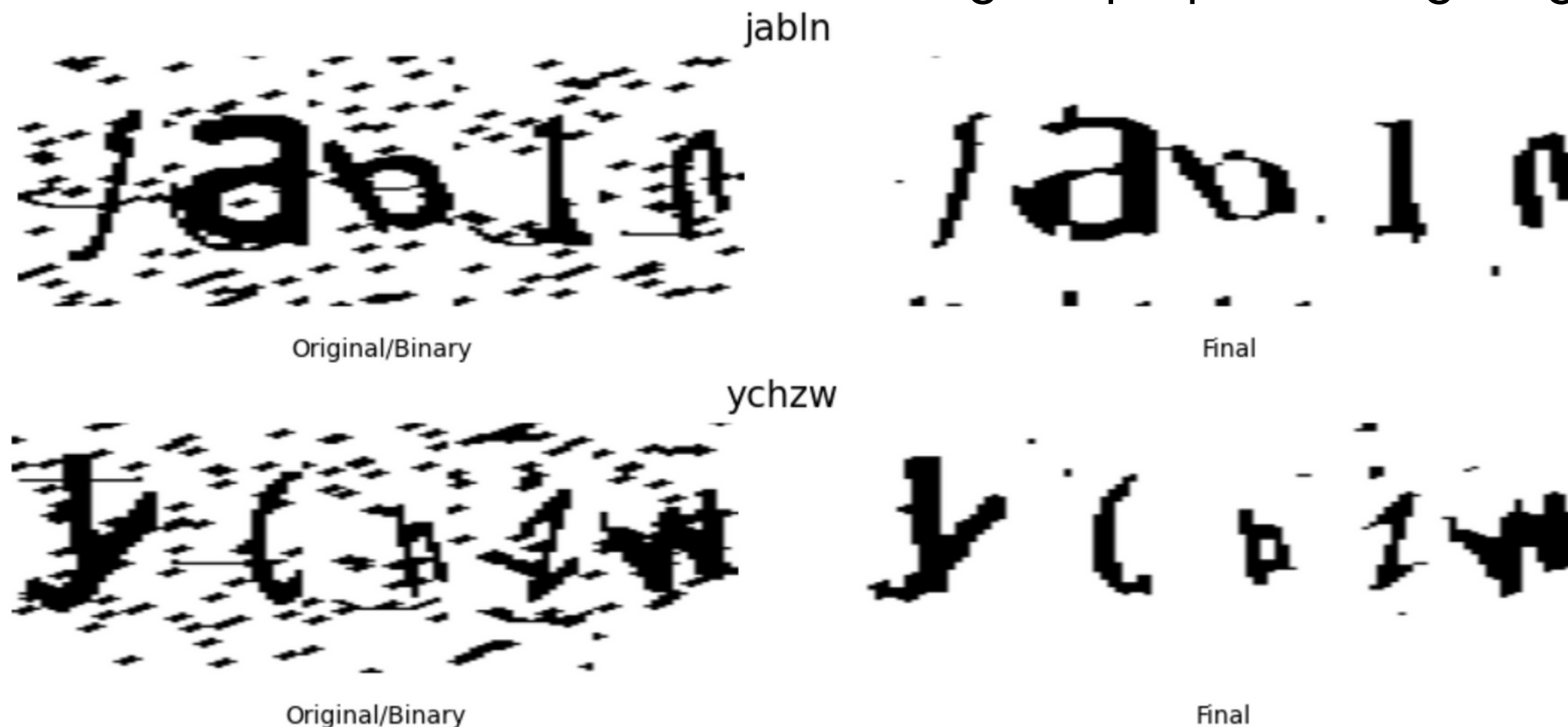
As can be seen from this bar graph, the characters b, f, h, i, j, l, o, t have low accuracy values.



The low accuracy values for these characters are due to unintended modifications that occurred during the preprocessing stage.



LOW ACCURACY



This misclassifications can be observed from the confusion matrix. For instance, a 'f' has been misclassified as a 'l' 9 times.

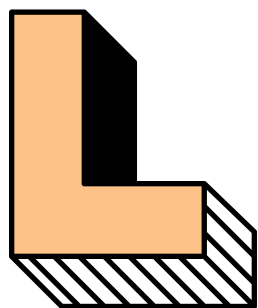


OTHER ALGORITHM TO IMPROVE ACCURACY: NEURAL NETWORKS

Neural networks, particularly deep learning models, have gained significant popularity in recent years due to their ability to learn complex patterns and relationships in data. Architectures like Convolutional Neural Networks (CNNs) for image data or Recurrent Neural Networks (RNNs) for sequential data can be powerful in improving accuracy for specific tasks.

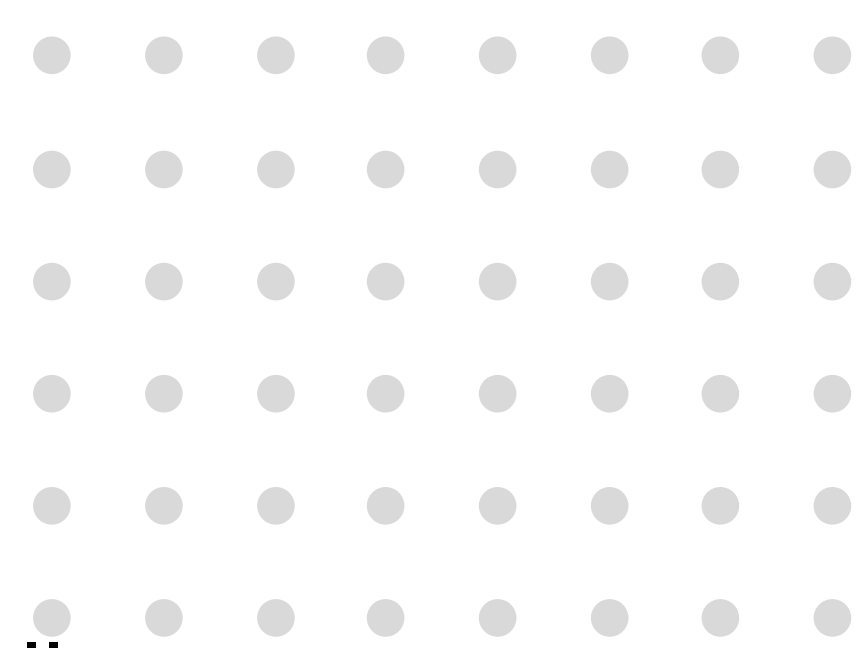
CNNs (Convolutional Neural Networks) are often considered better than Random Forest classifiers for tasks involving image classification because CNNs have a deeper understanding of the image features.

They are specifically designed for analyzing images. The advantage lies in their ability to automatically learn and extract relevant features from images. They use a technique called convolution, which involves sliding a small filter across the image and capturing local patterns. These learned features are then combined to make predictions about the image content.

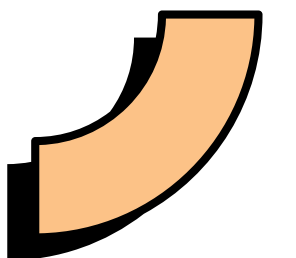




CHALLENGES FACED



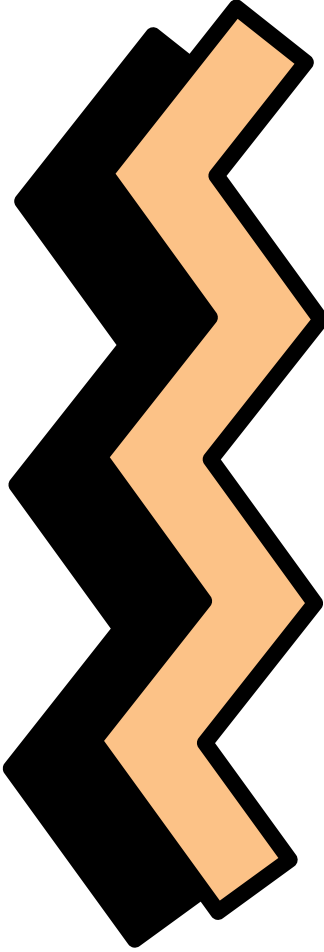
- We had used Google Colab for the first time and so initially faced some difficulties in using it.
- The program's long execution time lengthened the debugging and analysis process.
- Obtaining the best preprocessed image required significant time and experimentation
- Selecting optimal hyperparameters and the range of values was challenging and required extensive experimentation to achieve optimal performance.
- We researched some additional algorithms to improve accuracy and came across Neural Networks. But on running the code for it, we couldn't get the desired accuracy.
- Debugging the code was a challenging task.

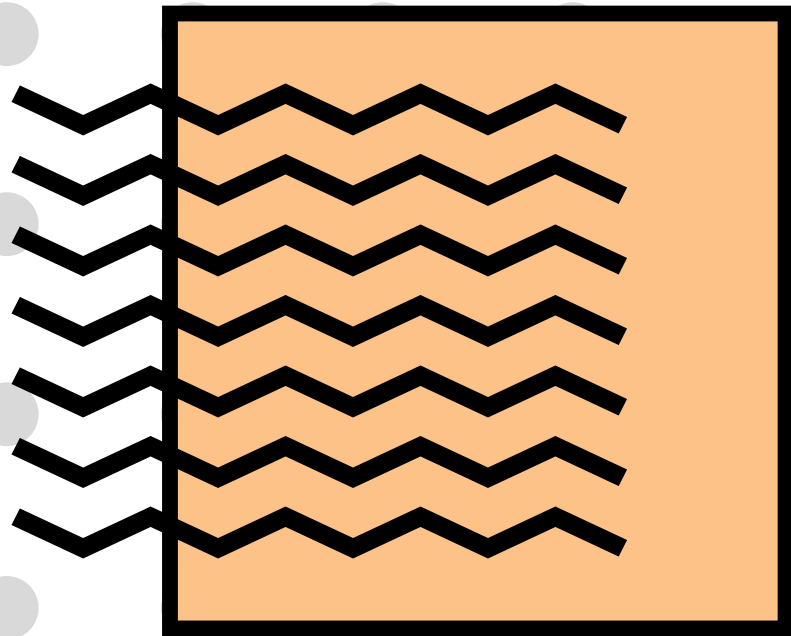




CONCLUSIONS



- 
- Random Forest model performed the best in terms of accuracy among the evaluated models.
 - Changes in preprocessing techniques had a significant impact on the choice of the best model (Random Forest or SVMs) and their respective performances.
 - Train-test splitting methods can influence model performance, and character-wise splitting (e.g., 80% of instances of each character in the training set) resulted in increased accuracy.
 - Accuracy alone may not be the most reliable metric for evaluating classifier performance, especially in imbalanced datasets. Precision, recall, and F1 score provide a more comprehensive assessment.
 - Precision, recall, and F1 scores were calculated to evaluate the performance of the classifiers.
 - The low accuracy for certain characters was attributed to unintended modifications during the preprocessing stage, affecting the training and classification of those characters.
 - Convolutional Neural Networks (CNNs) have the ability to achieve high accuracy values in various tasks, including image classification.



**THANKS FOR
WATCHING**

