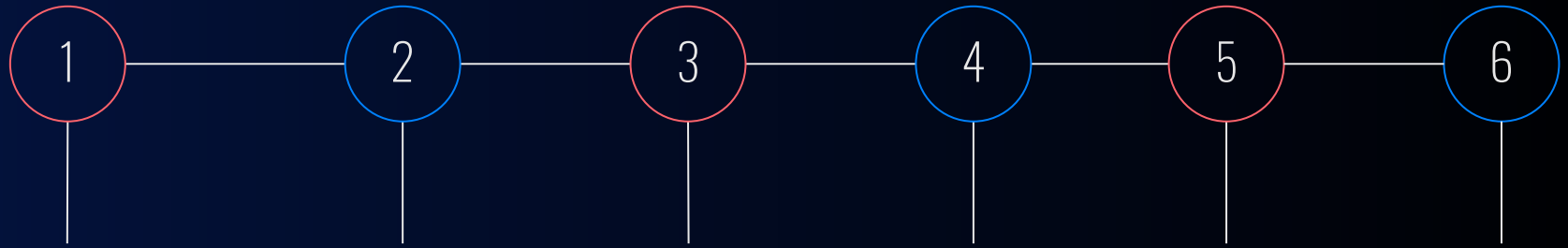
SignLingo

Developing a State-of-the-Art Interpreter Model for Sign Language Communication

Allen Lau, Shubham Khandale, Sumaiya Uddin

Agenda



Project Overview

What is the motivation? Why should you care?

Data & EDA

What is the dataset for model training?

Methodology

What methods and models did we use?

Evaluation

How do our models perform?

Demo

What does the Sign Language Interpreter look like?

Conclusions

What are the main takeaways?

01. Project Overview

~2 Million

Americans are classified as deaf

https://www.startasl.com/american-sign-language/#:~:text=American%20Sign%20Language%20(ASL)%20is,in%20the%20US%20and%20Canada.

500,000

people use ASL as their primary language in the U.S. and Canada

https://www.startasl.com/american-sign-language/#:~:text=American%20Sign%20Language%20(ASL)%20is,in%20the%20US%20and%20Canada.

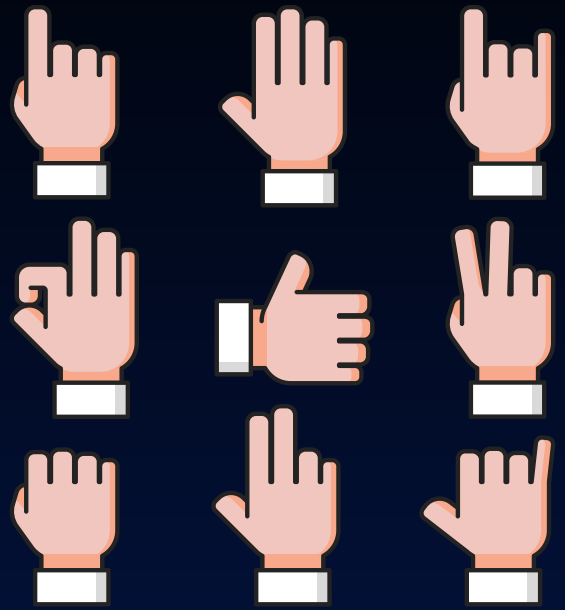
3

Is the ranking of ASL in the ranking of the most commonly used languages in the U.S.

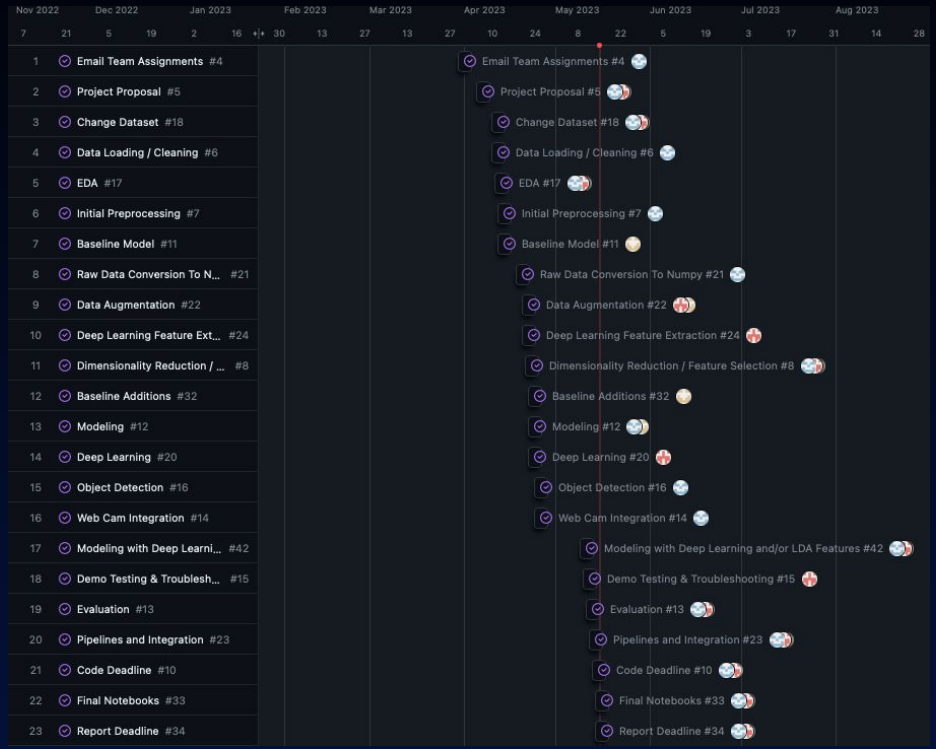
https://www.startasl.com/american-sign-language/#:~:text=American%20Sign%20Language%20(ASL)%20is,in%20the%20US%20and%20Canada.

Why SignLingo?

- Effective communication is a vital part of society
- Bridge the gap of communication between those who rely on sign language and those who do not know sign language
- Potential Applications:
  - Ease of communication in settings like schools and government buildings
  - Reduce potential hurdles for the hearing impaired or the deaf
  - Creation of an easily packaged model that can be deployed



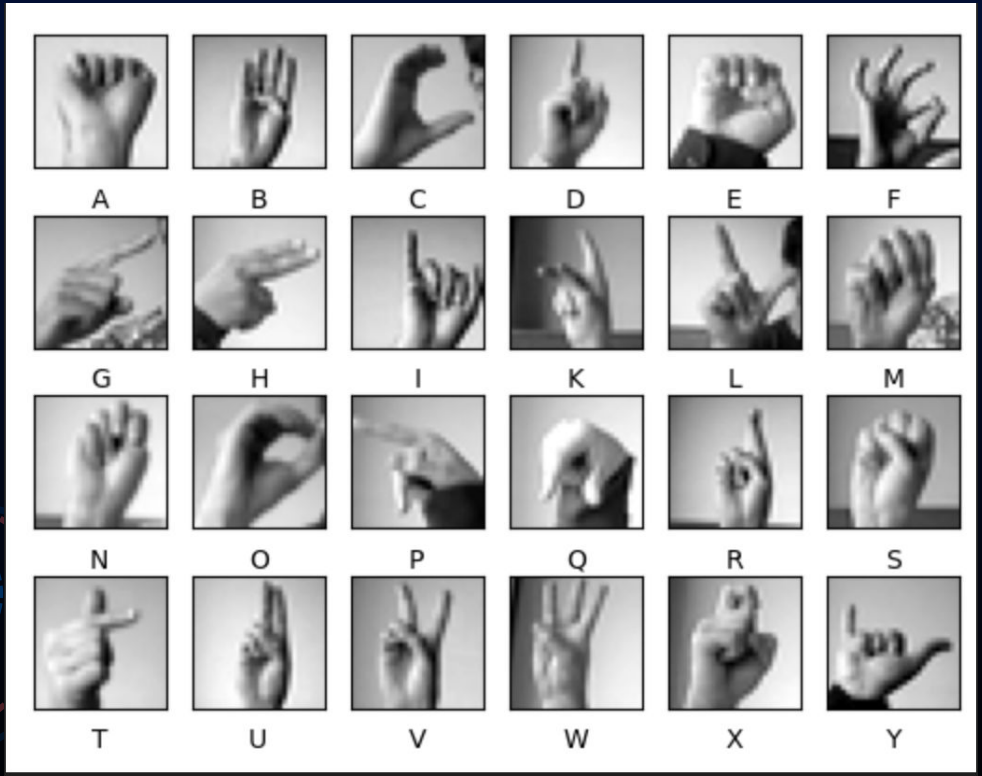
Github Repository



https://github.com/DataScienceAndEngineering/machine-learning-dse-i210-final-project-signlanguageclassification.git

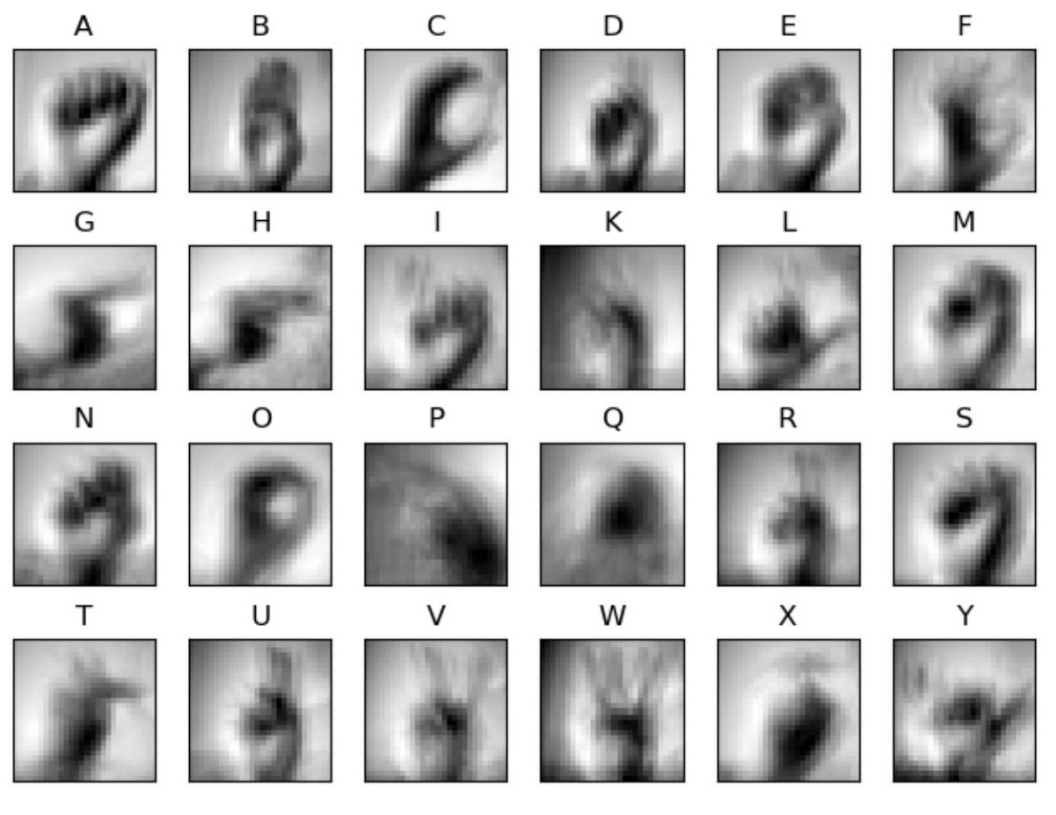
02. Data & EDA

Sign Language Image Dataset

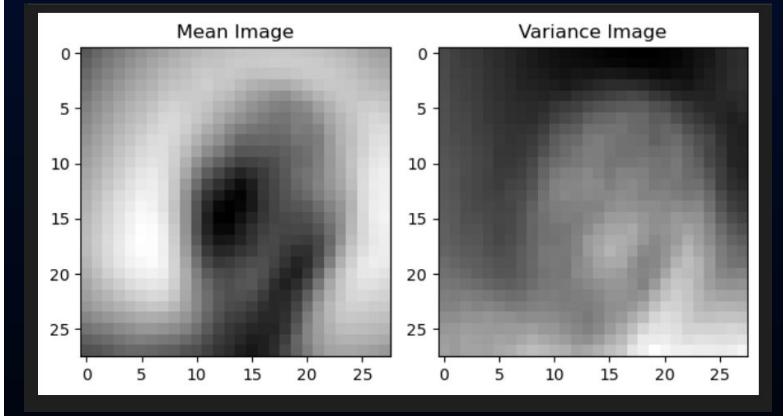


- 28x28 Grayscale Images
  - Training Data: 27,455
  - Testing Data: 7,172
- Labels
  - 24 classes of letters (excluding J and Z, which require motion)
- Features
  - Pixel intensity values
  - 0 - 255

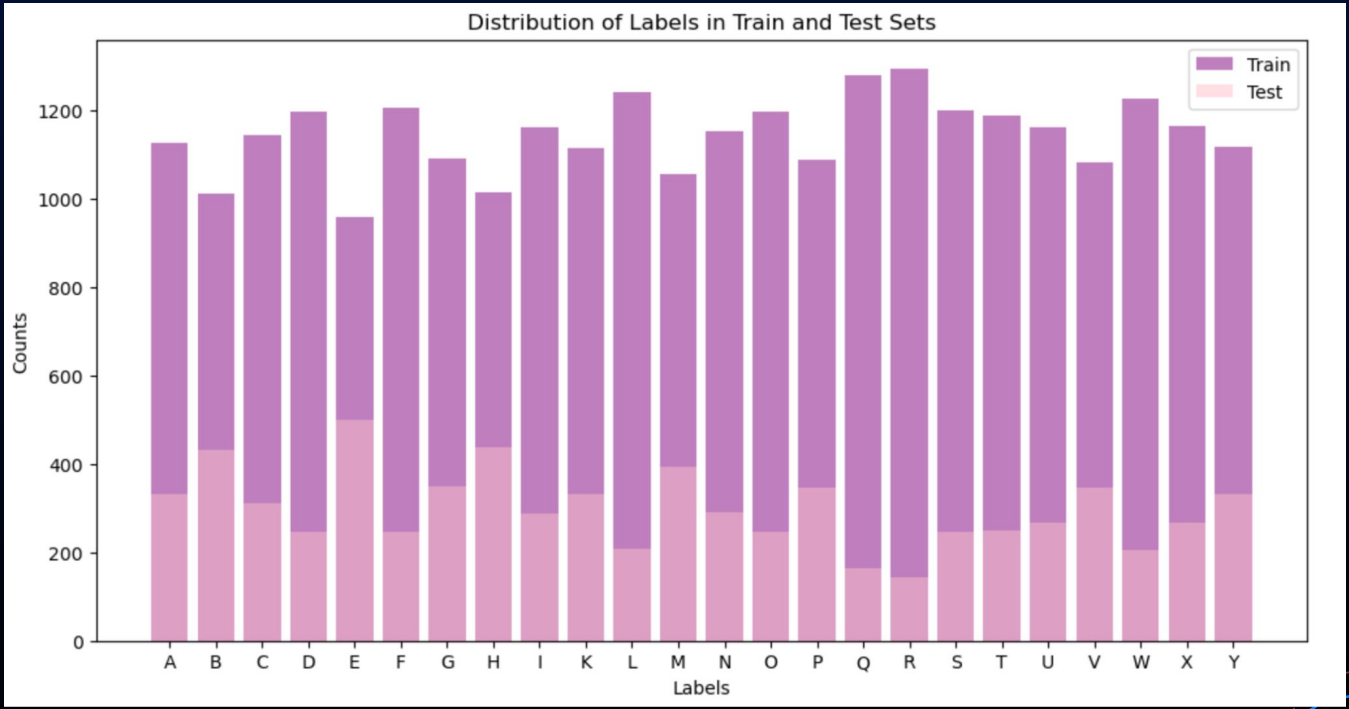
Mean & Variance Images



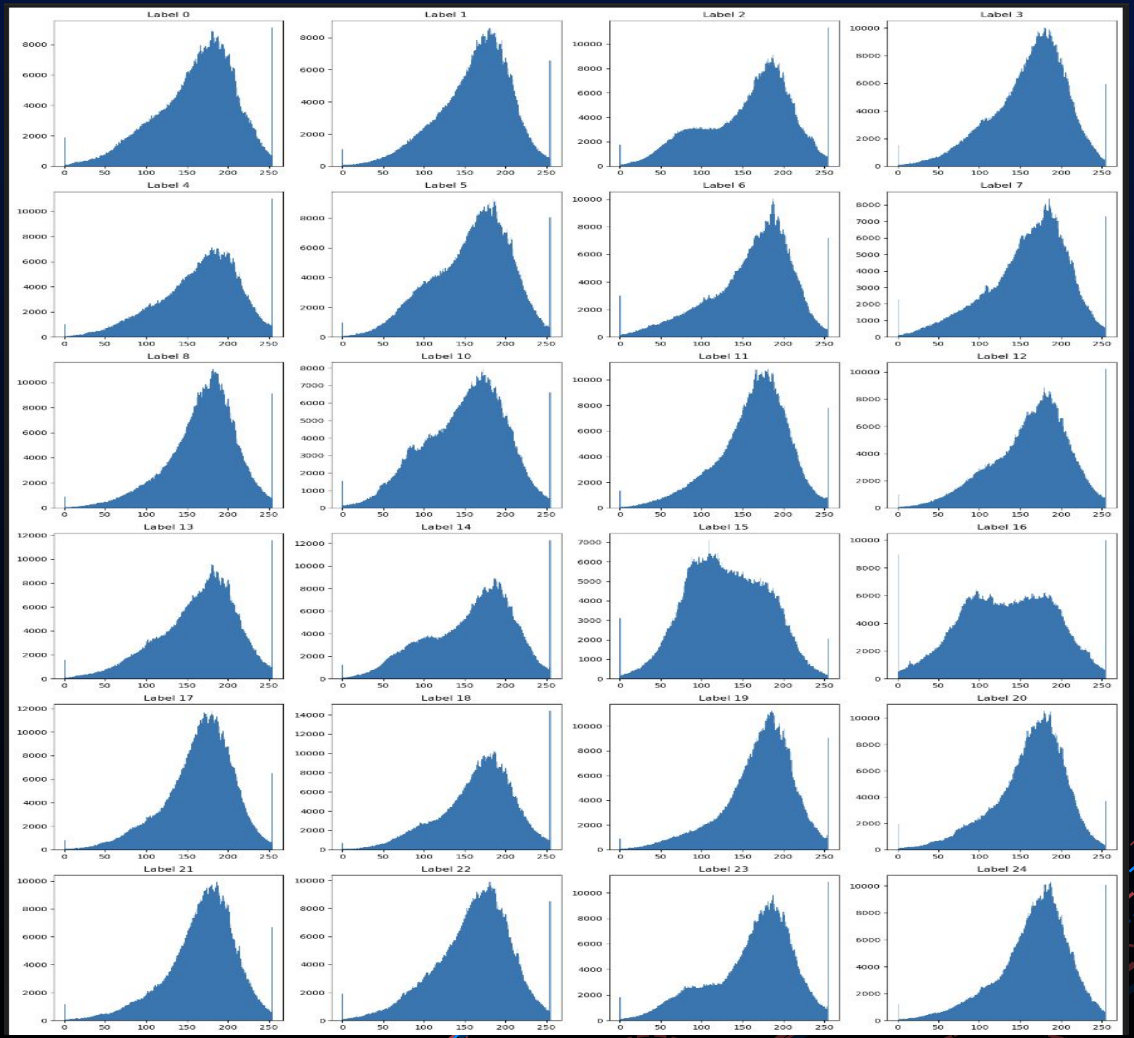
Overall Mean & Variance Images



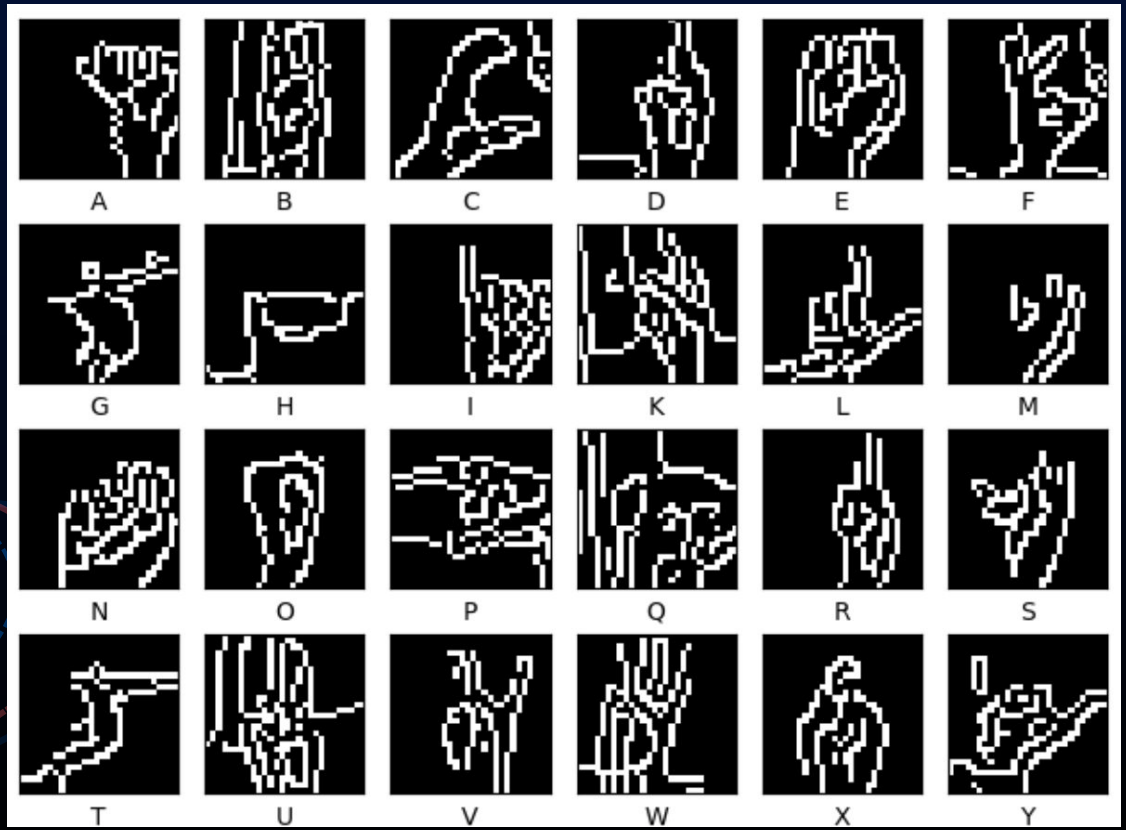
Class Distributions



Distributions of Pixel Intensities per Label



Canny Edge Detection



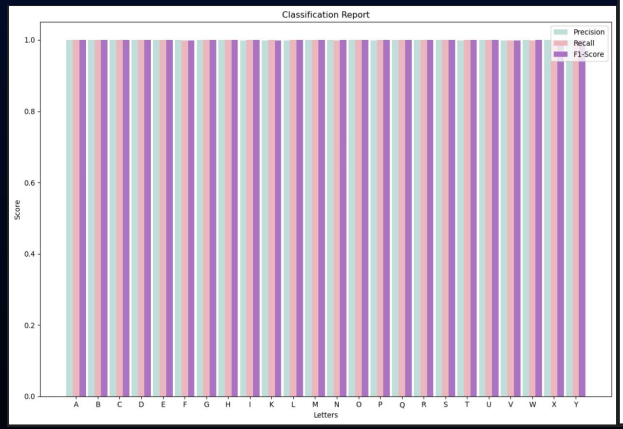
- Sigma parameter to control the threshold range
- Lower and upper threshold based on median pixel value
- 3x3 kernel to connect broken lines and close gaps

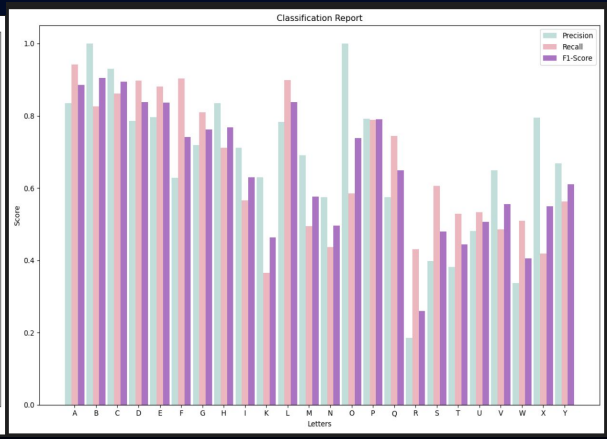
https://www.reddit.com/r/machinelearningmemes/comments/f2m7yc/another_overfitting_meme/

Initial Model Results

- Models Trained
  - Naive Bayes
  - Logistic Regression
  - Random Forest
  - Support Vector Machine
- Hyperparameter Tuning
  - Randomized Grid Search CV
- Overfit!

Logistic Regression Results





Training Accuracy 1.00 and Testing Accuracy 0.67

Best Hyperparameters:

- C : 3.4647045830997407
- Max_inter : 3171
- Penalty : 12
- Solver : liblinear
- Warm_start : False

Strategies to Combat Overfitting

Regularization

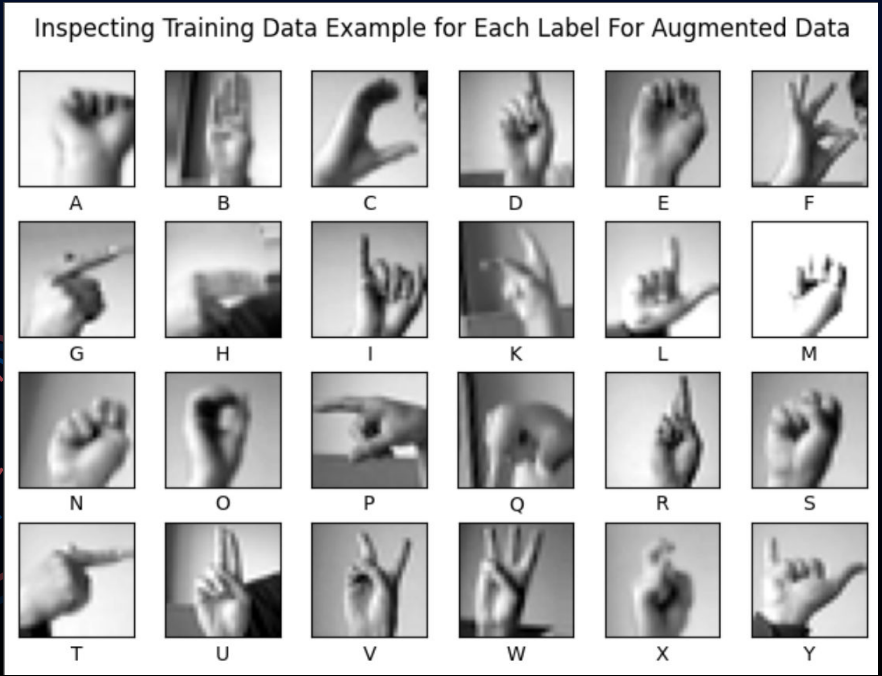
Data Augmentation

Dimensionality Reduction

Data Augmentation

Hyperparameters:

- Rotation_range : 10
- Zoom_range : 0.1
- Width_shift_range : 0.1
- Height_shift_range : 0.1
- Shear_range : 0.1
- Brightness_range : [0.5,1.5]
- Fill_mode : 'nearest'

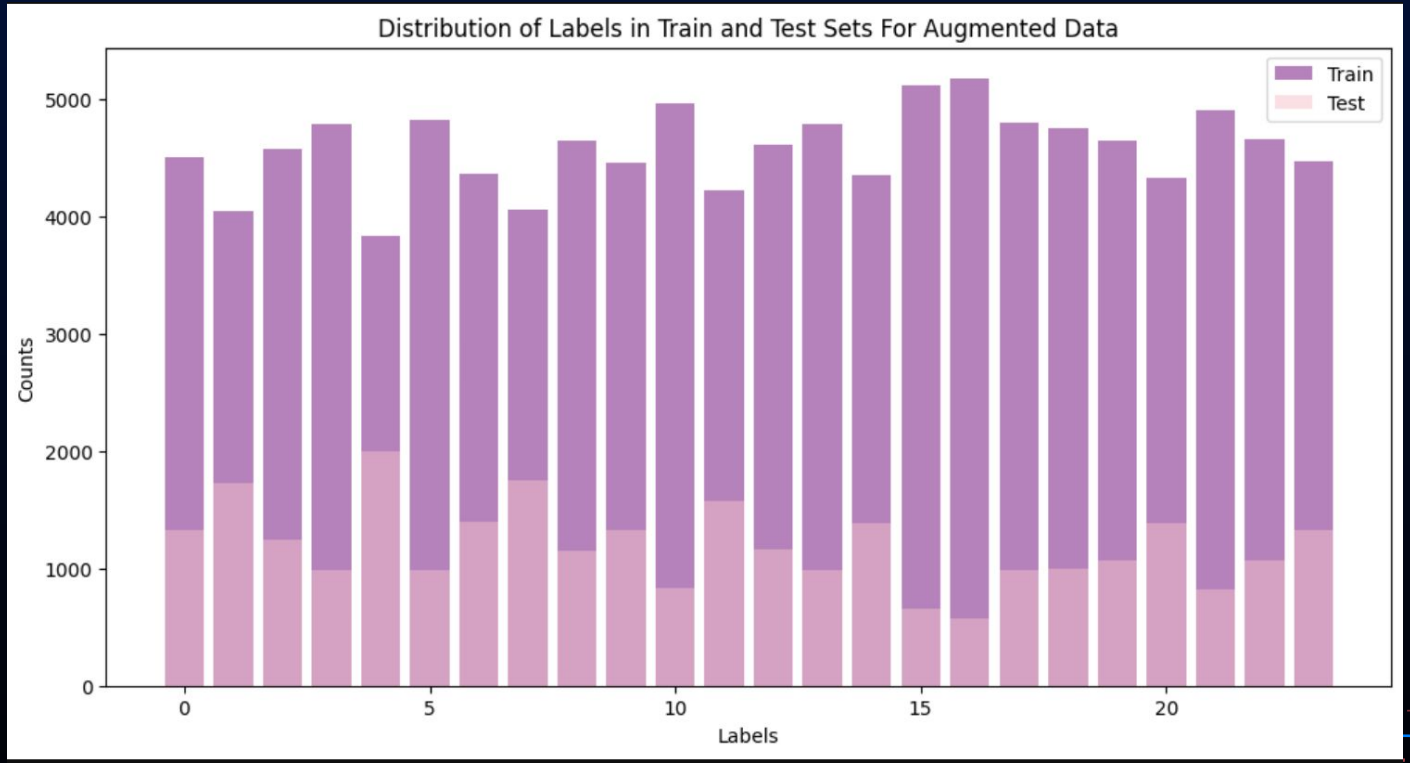


Data Augmentation: Counts

Training Data: 27,455
Testing Data: 7,172

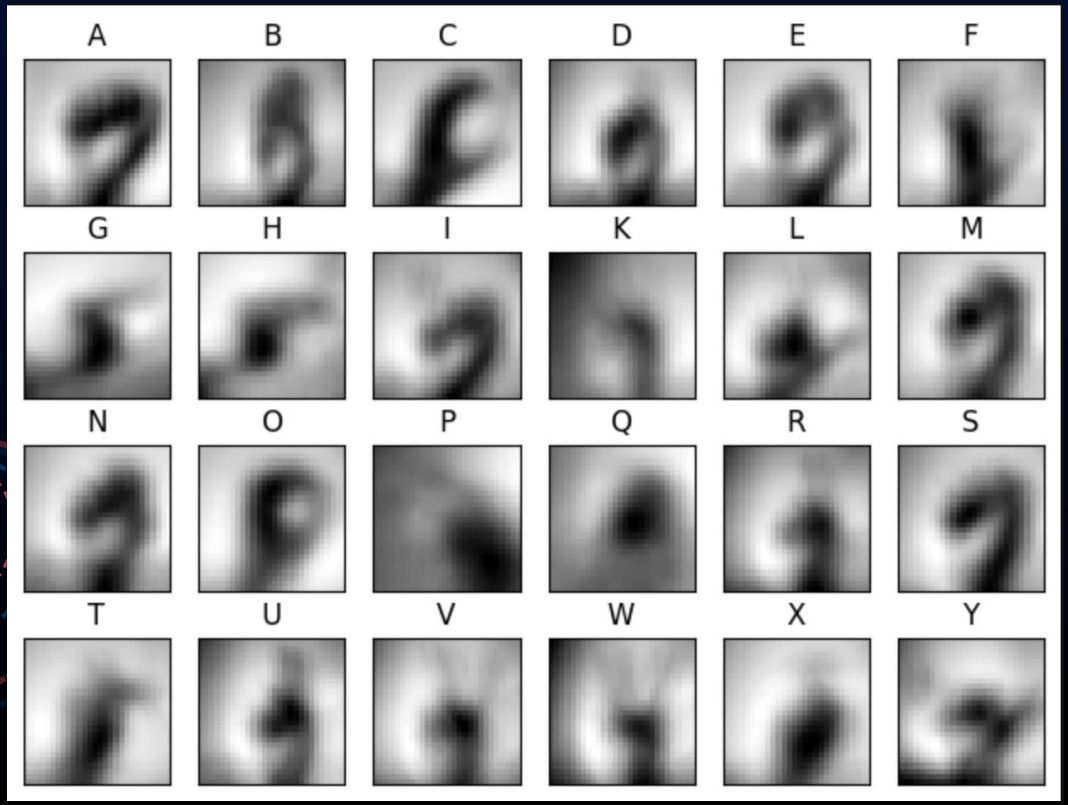
→

Training Data: 109,820
Testing Data: 28,688

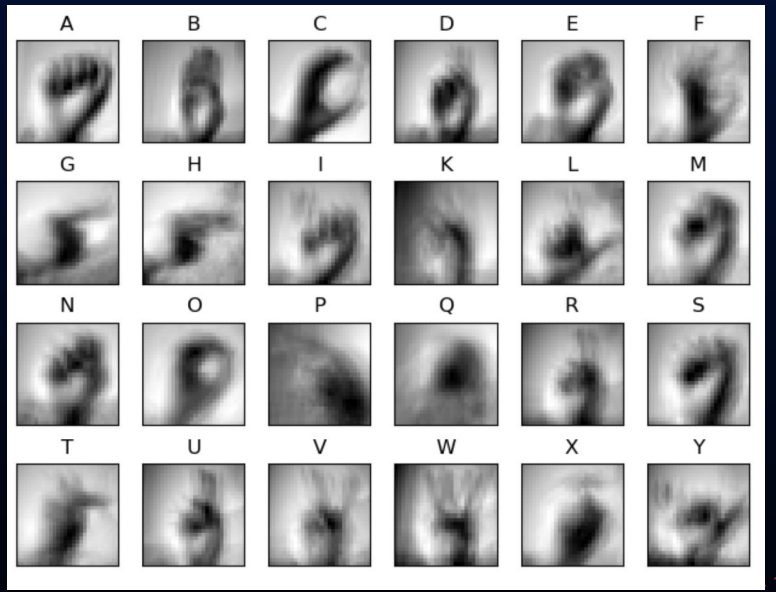


Data Augmentation: Mean Image

Augmented Data



Original Data

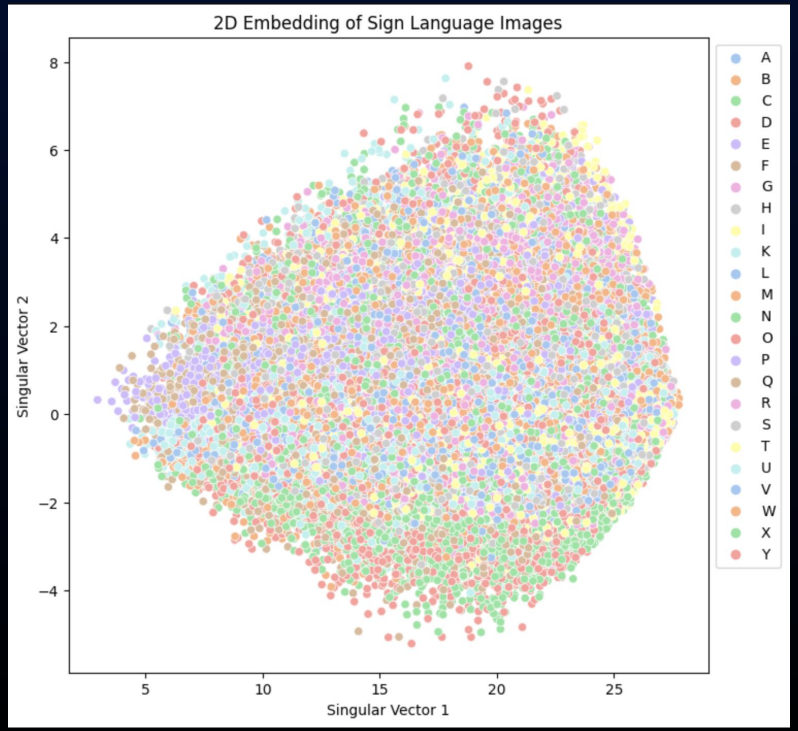


03. Methodology

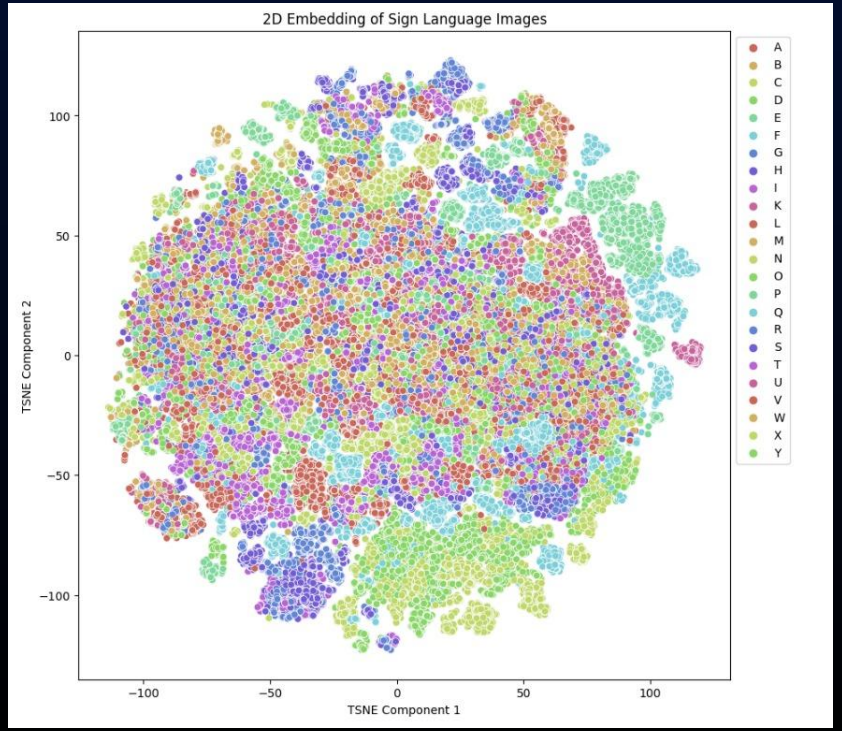
Dimensionality Reduction

Dimensionality Reduction: SVD & TSNE

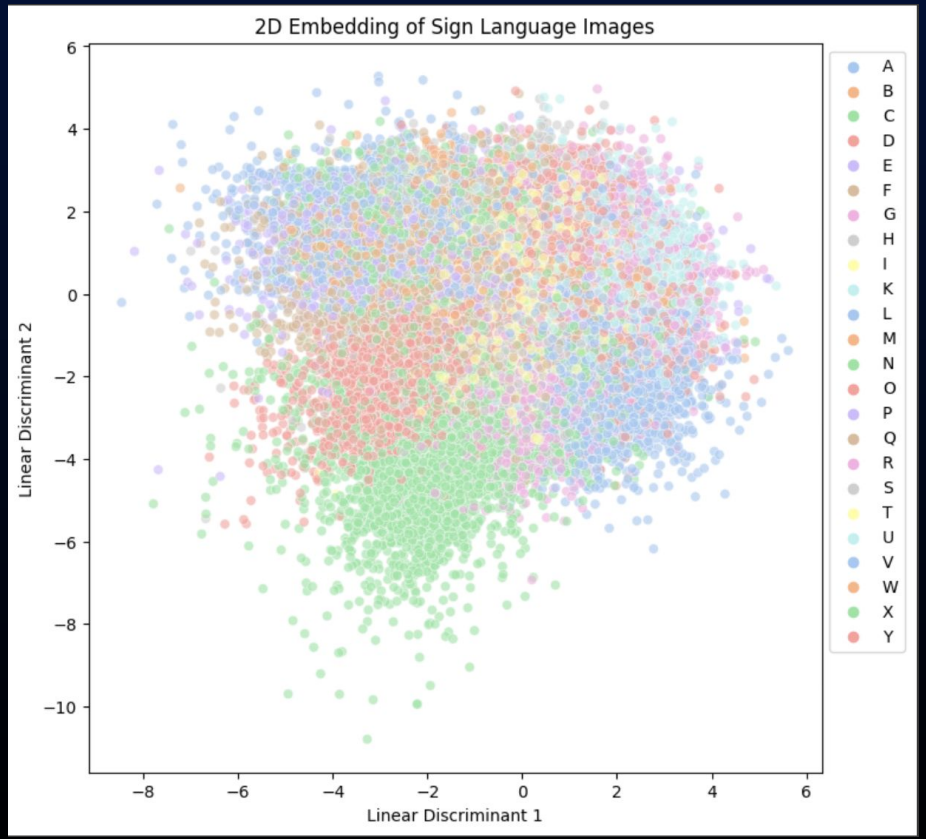
Singular Value Decomposition

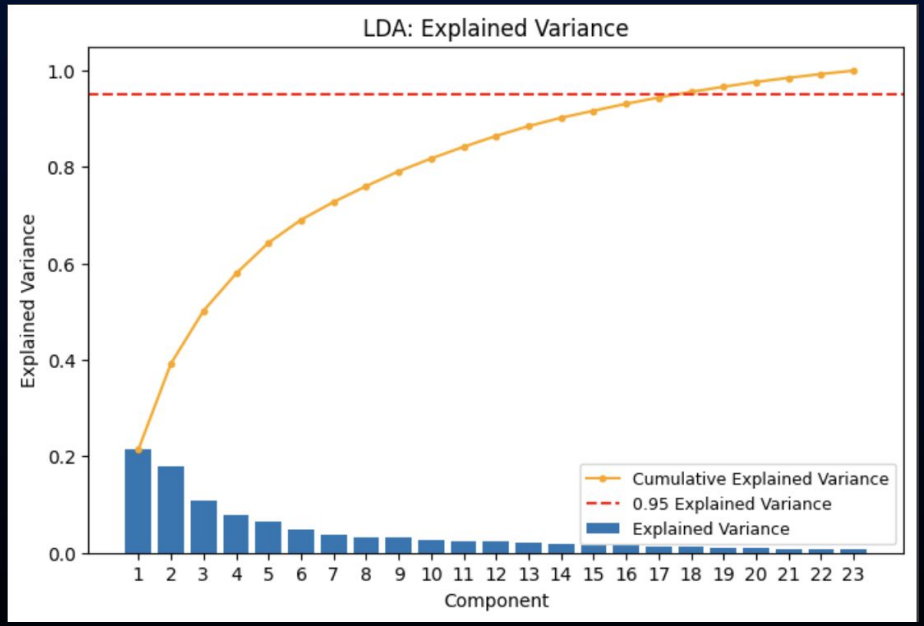


T-distributed Stochastic Neighborhood Embedding

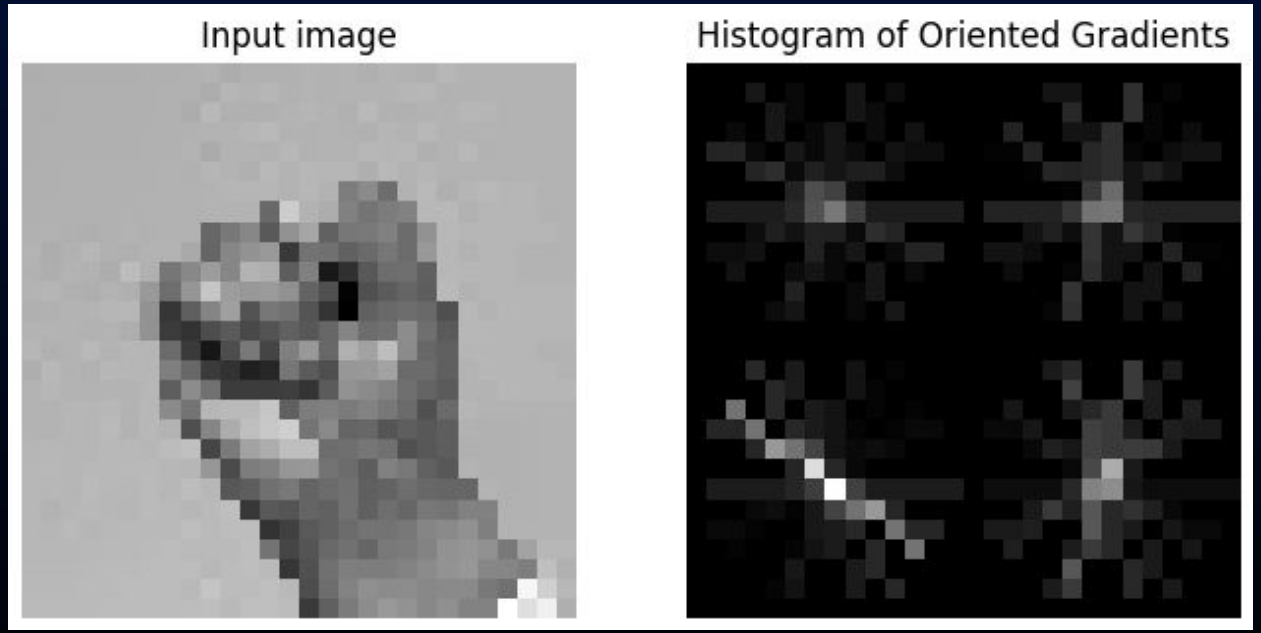


Dimensionality Reduction: LDA





Feature Engineering: Histogram of Oriented Gradients

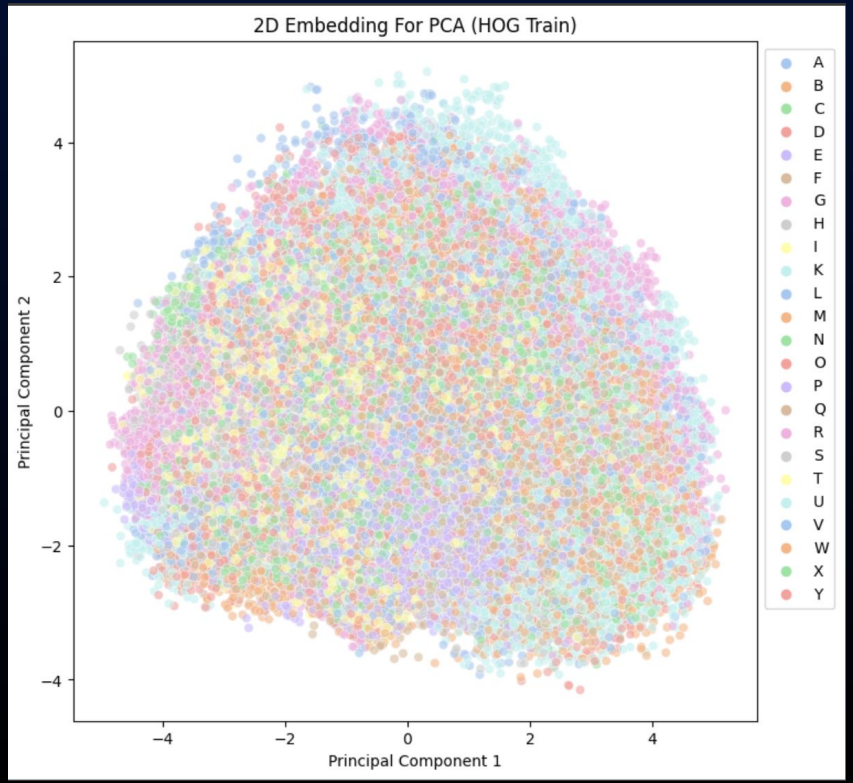


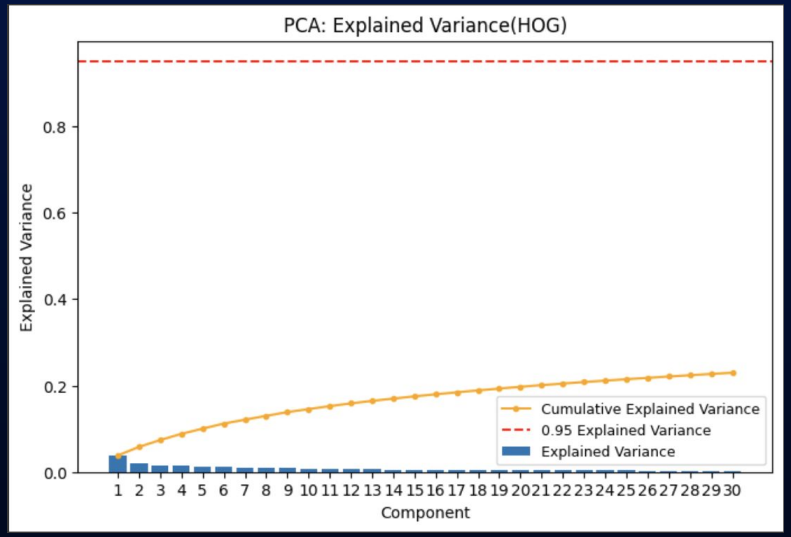
784 features → 1764 features

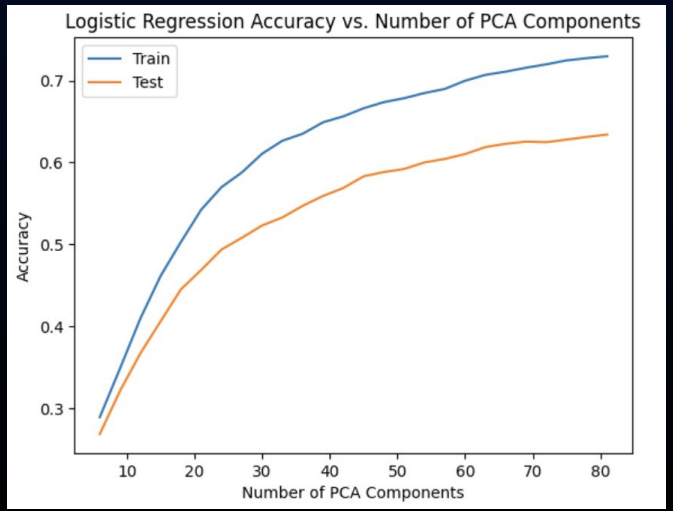
Parameters:

- Orientations : 9
- pixels_per_cell : (14,14)
- Cells_per_block : (2,2)
- Block_norm : 'L2-Hys'

Dimensionality Reduction: PCA on HOG Features

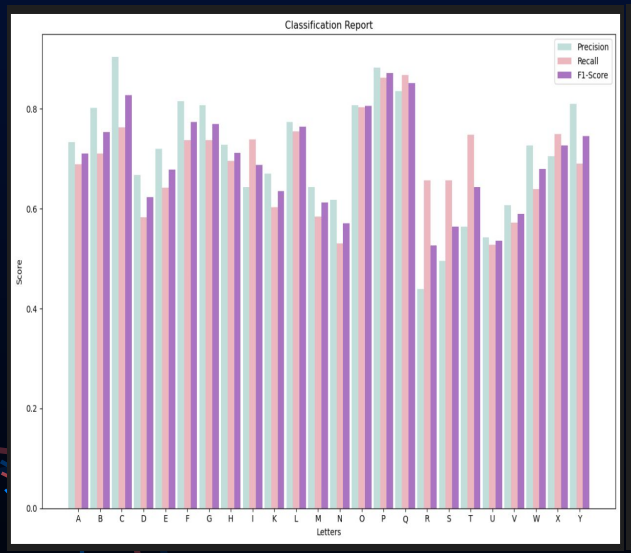


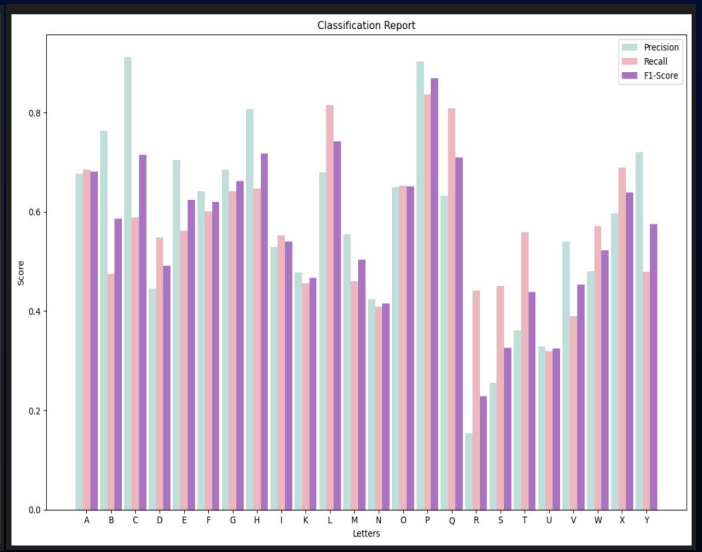




Naive Bayes

Naive Bayes: Evaluation Metrics

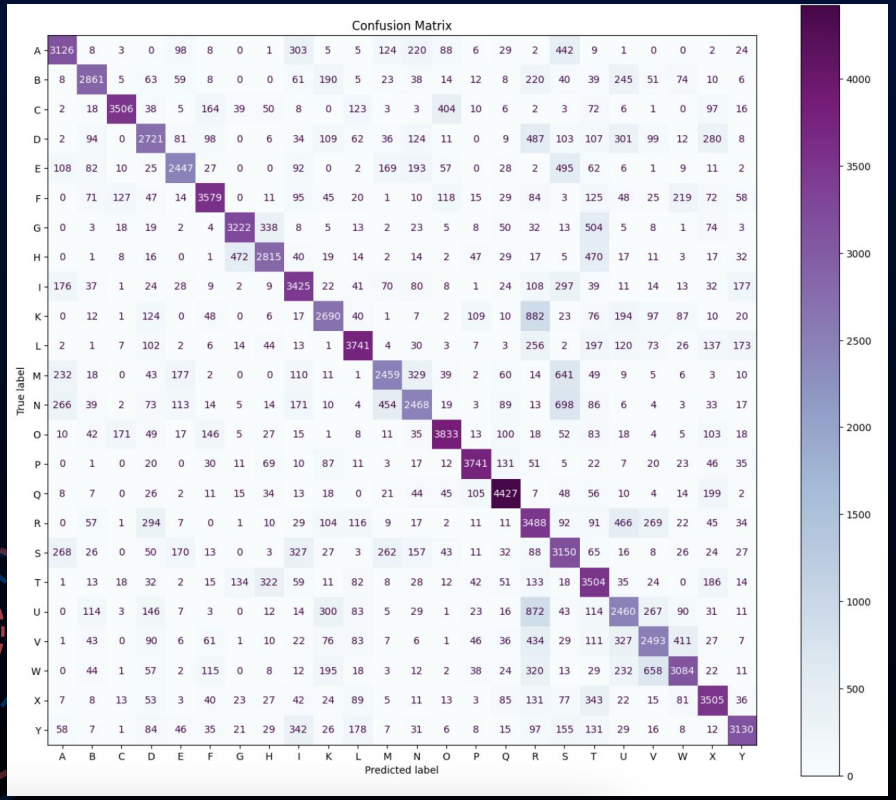


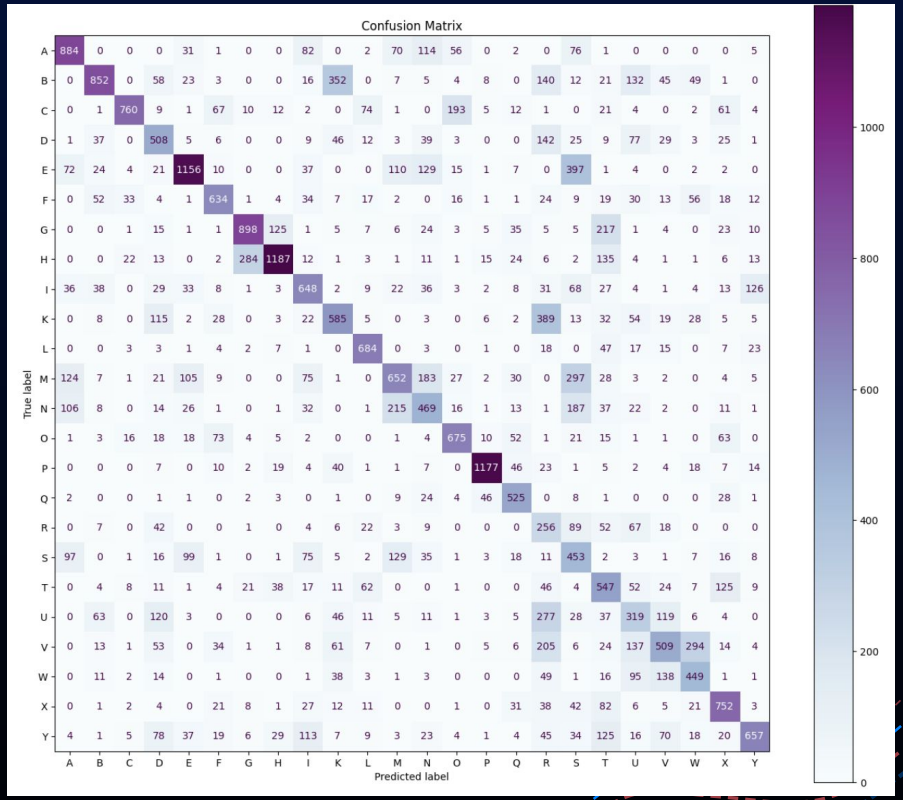


Evaluation Summary

- Accuracy: For train, accuracy is 69% and, and test is 57%
- Precision: A-Y labels range from 0.45-0.90 for train and 0.15-0.91 for test. P & C have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.54-0.87 for train and 0.30-0.85 for test, P & H have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.6778 for train, indicating a moderate level of agreement between predicted and true labels. For test, it is approximately 0.548.
- Letter R an S had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classes
- Cohen's Kappa is a statistical measure of inter-rater agreement for categorical classifications, which is moderate for this model

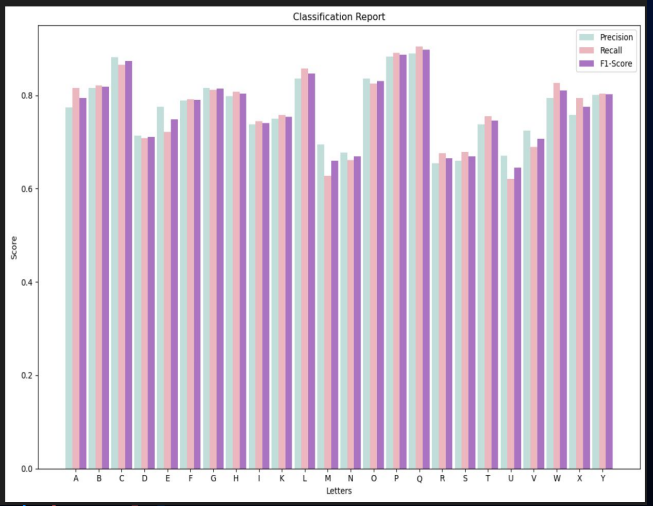
Naive Bayes: Confusion Matrix

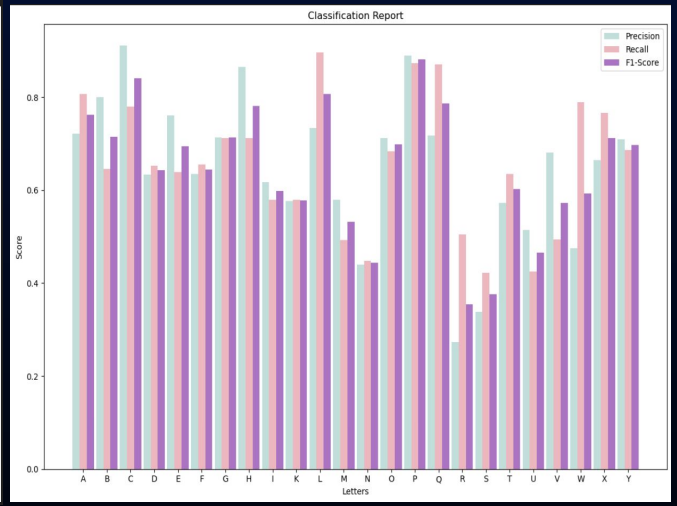




Logistic Regression

Logistic Regression: Evaluation Metrics





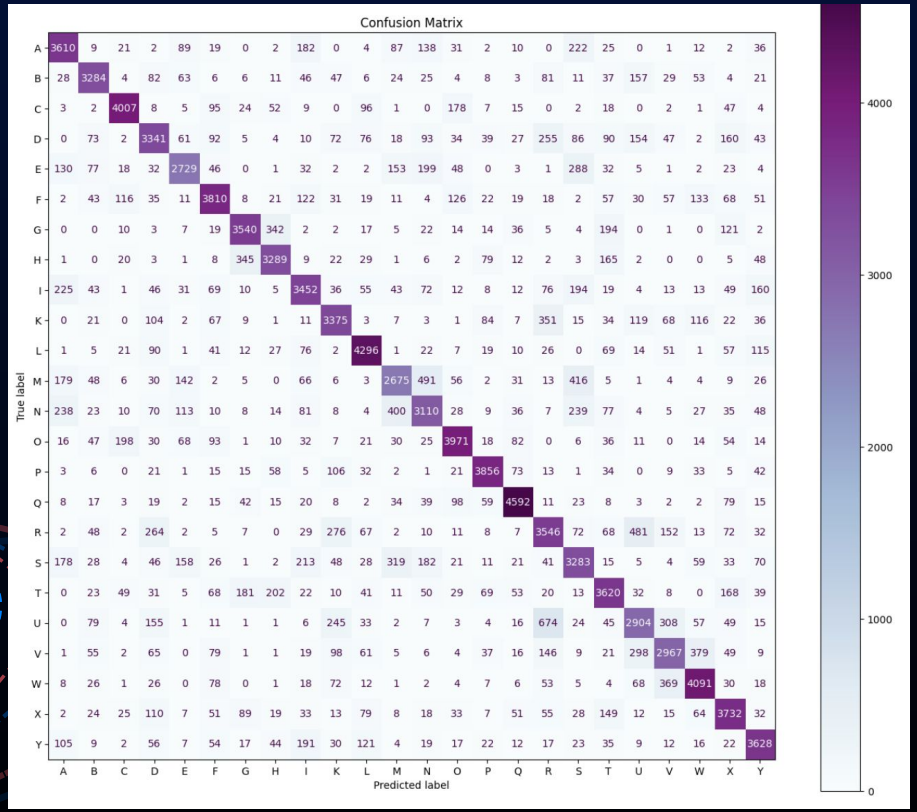
Best Hyperparameters:

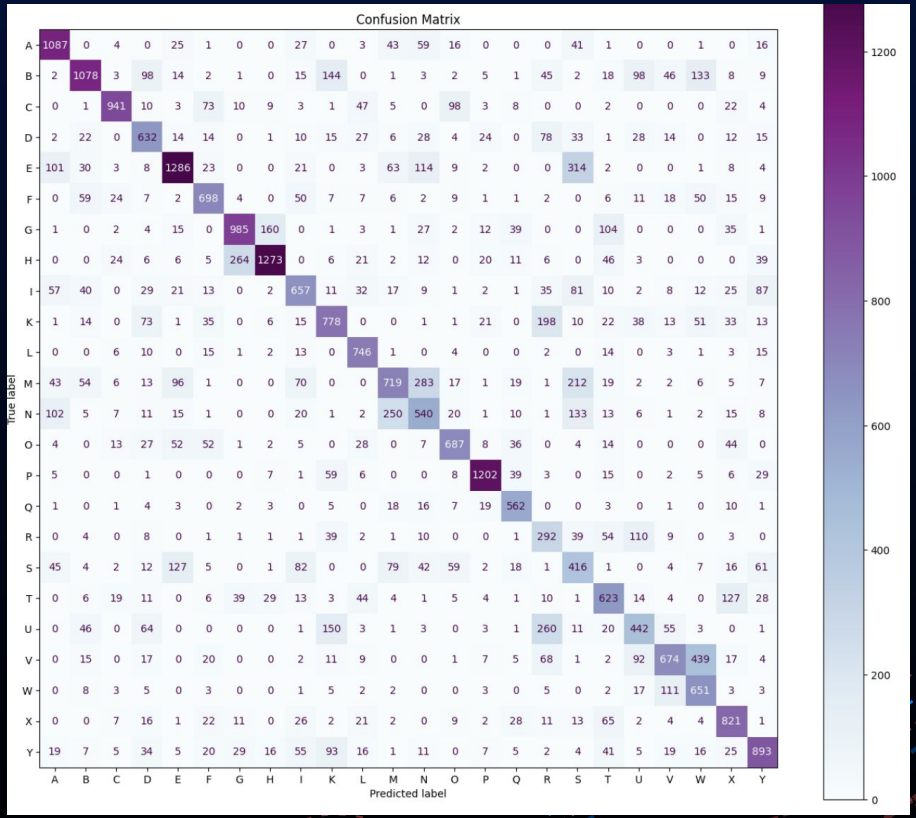
- C - 0.22564631610840102
- Max_inter : 2391
- Penalty : “12”
- Solver : “newton-cg”
- Warm_start : False

Evaluation Summary

- Accuracy: For train, accuracy is 77% and, and test is 65%
- Precision: A-Y labels range from 0.66-0.89 for train and 0.29-0.89 for test. P, & C have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.63-0.90 for train and 0.42-0.89 for test, P, C & H have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.761 for train, indicating a little higher level of agreement between predicted and true labels. For test, it is approximately 0.636 which moderate level.
- Letter R an S had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classe

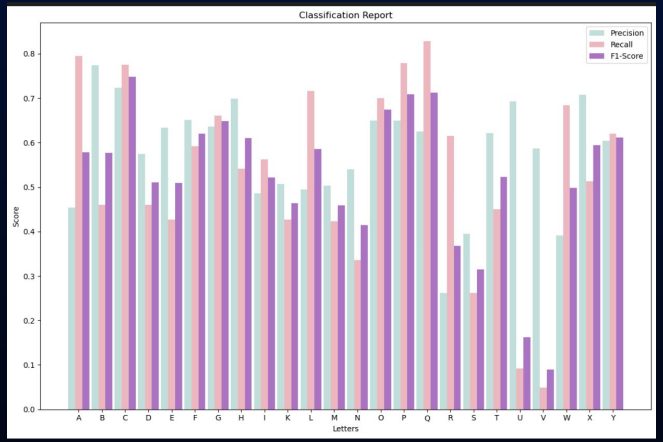
Logistic Regression: Confusion Matrix

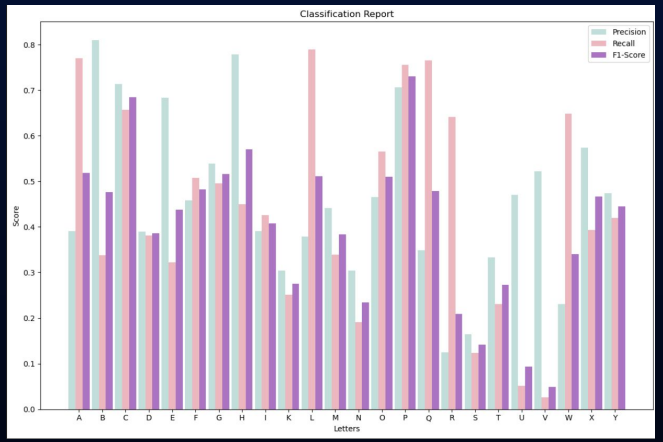




Random Forest

Random Forest: Evaluation Metrics





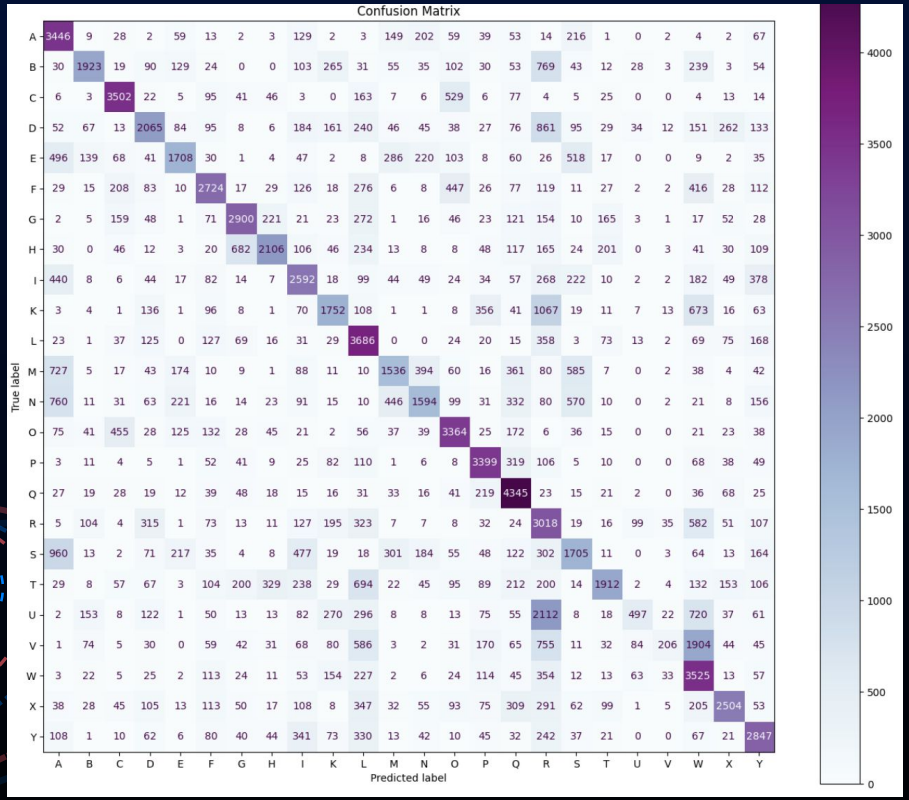
Best Hyperparameters:

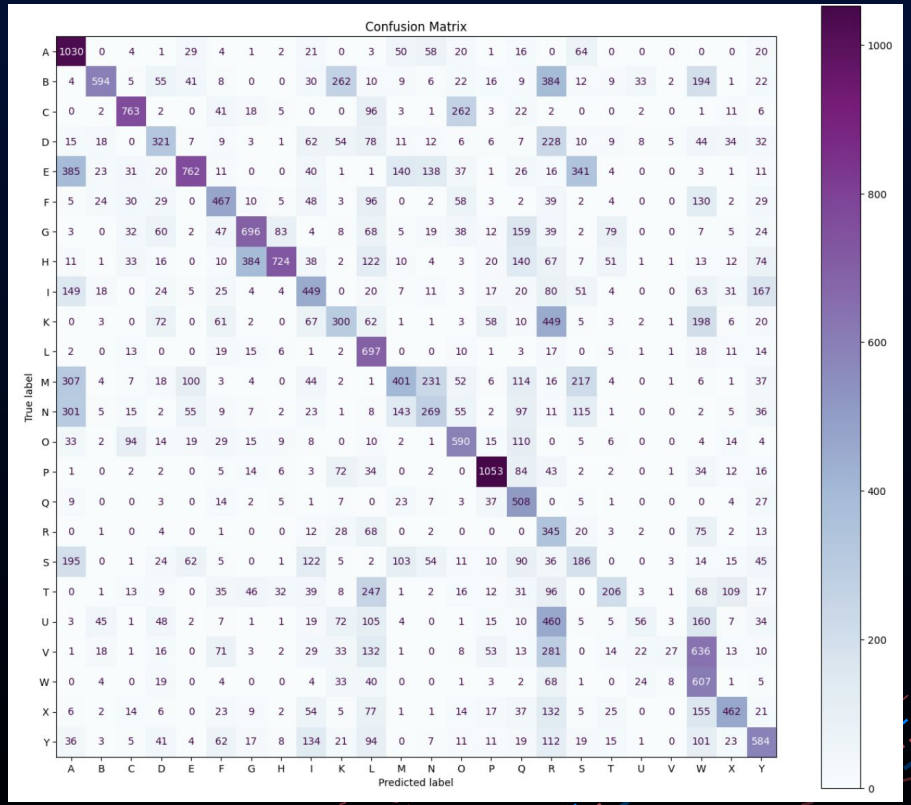
- N_estimators = 20
- Mmin_sample_split = 10
- Min_sample_leaf = 5
- Max_features = 5
- Max_depth = 5

Evaluation Summary

- Accuracy: For train, accuracy is 54% and, and test is 42%
- Precision: A-Y labels range from 0.27-0.74 for train and 0.12-0.77 for test. P, B & C have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.54-0.87 for train and 0.30-0.85 for test, P & H have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.518 for train, indicating a little lower level of agreement between predicted and true labels. For test, it is approximately 0.402.
- Letter R an S had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classes

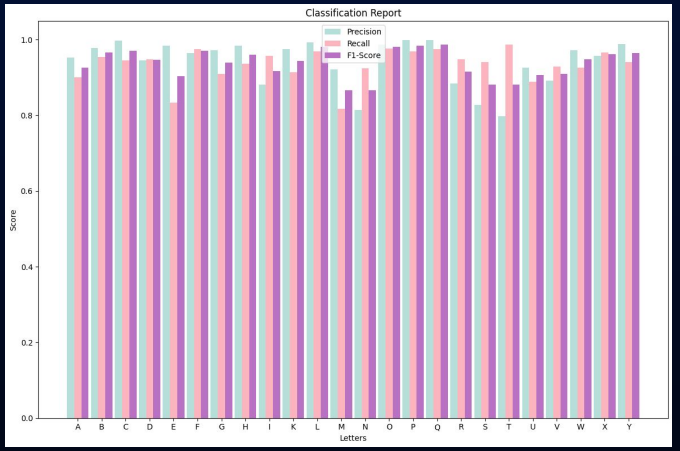
Random Forest: Confusion Matrix

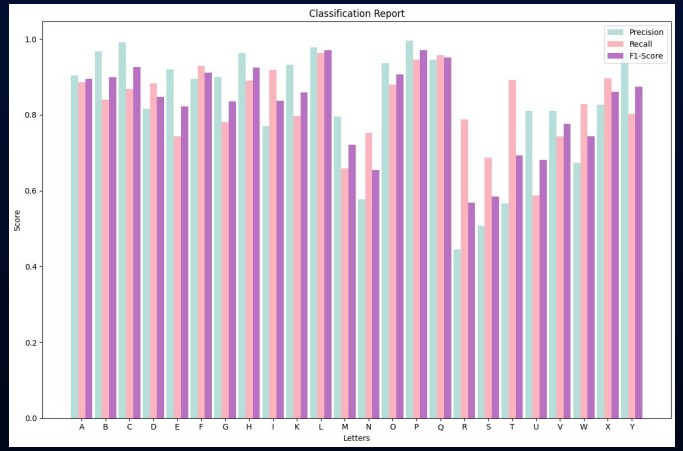




Support Vector Machine

SVM: Evaluation Metrics





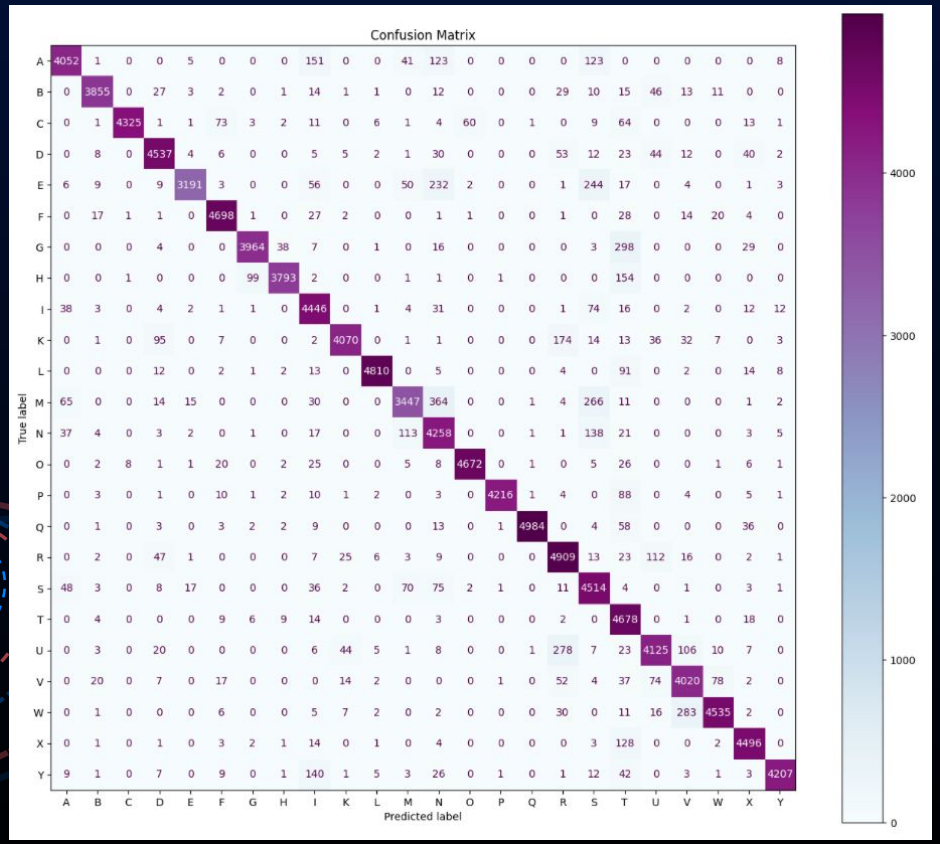
Best Hyperparameters:

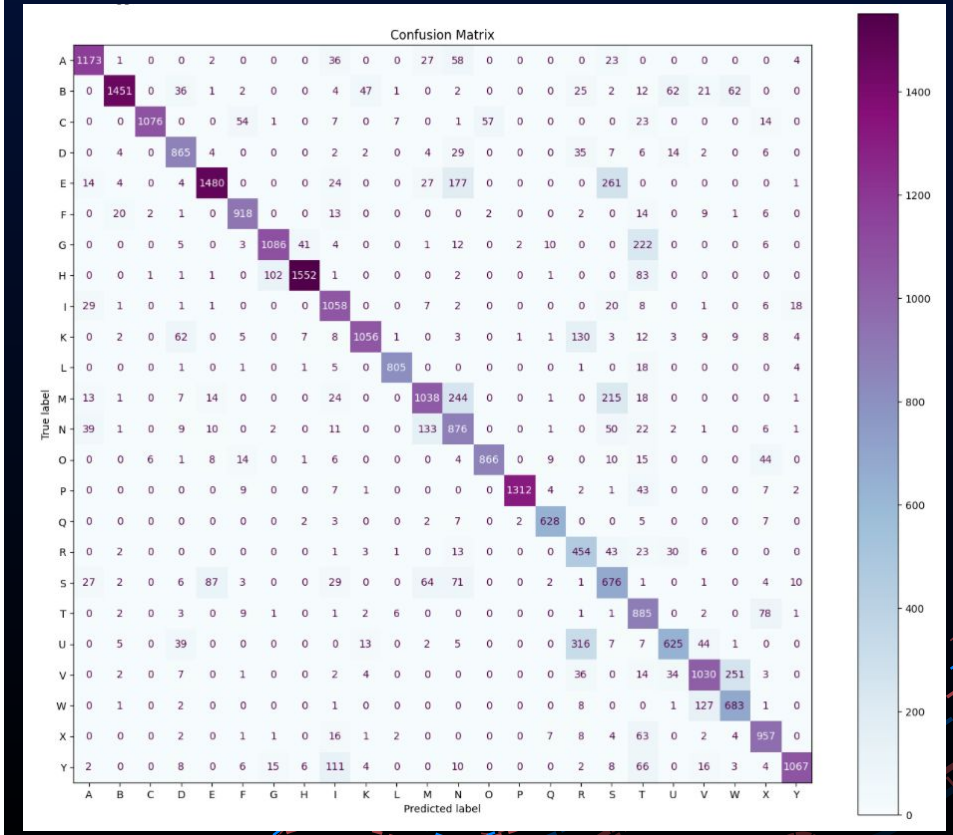
- Kernel : 'poly'
- Gamma : 'auto'
- C : 0.1

Evaluation Summary

- Accuracy: For train, accuracy is 94% and, and test is 82%
- Precision: A-Y labels range from 0.81-1.00 for train and 0.44-1.00 for test. P & C have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.82-0.99 for train and 0.74-0.96for test, L, Q & P have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.9334 for train, indicating a higher level of agreement between predicted and true labels. For test, it is approximately 0.8159.
- Letter R an S had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classes

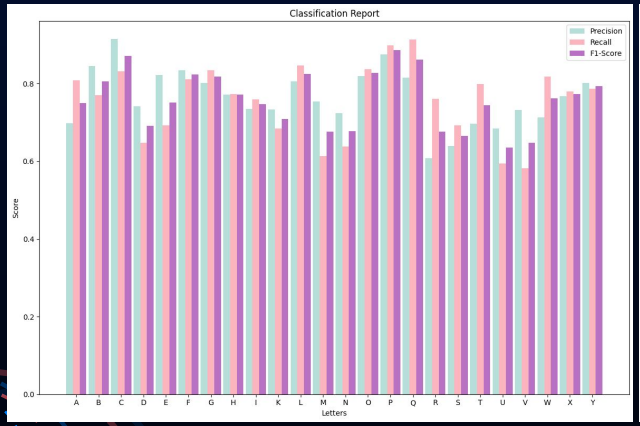
SVM: Confusion Matrix

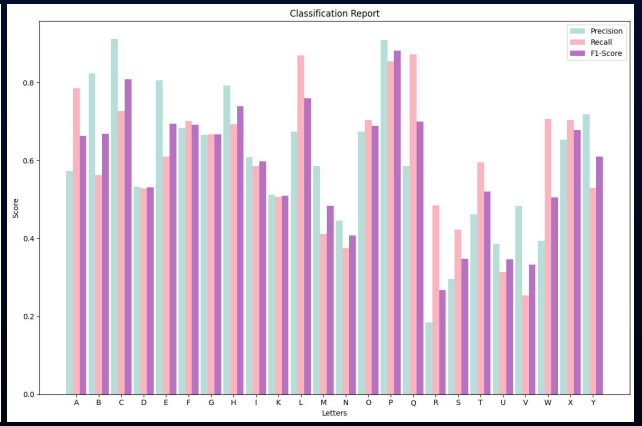




XGBoost

XGBoost: Evaluation Metrics





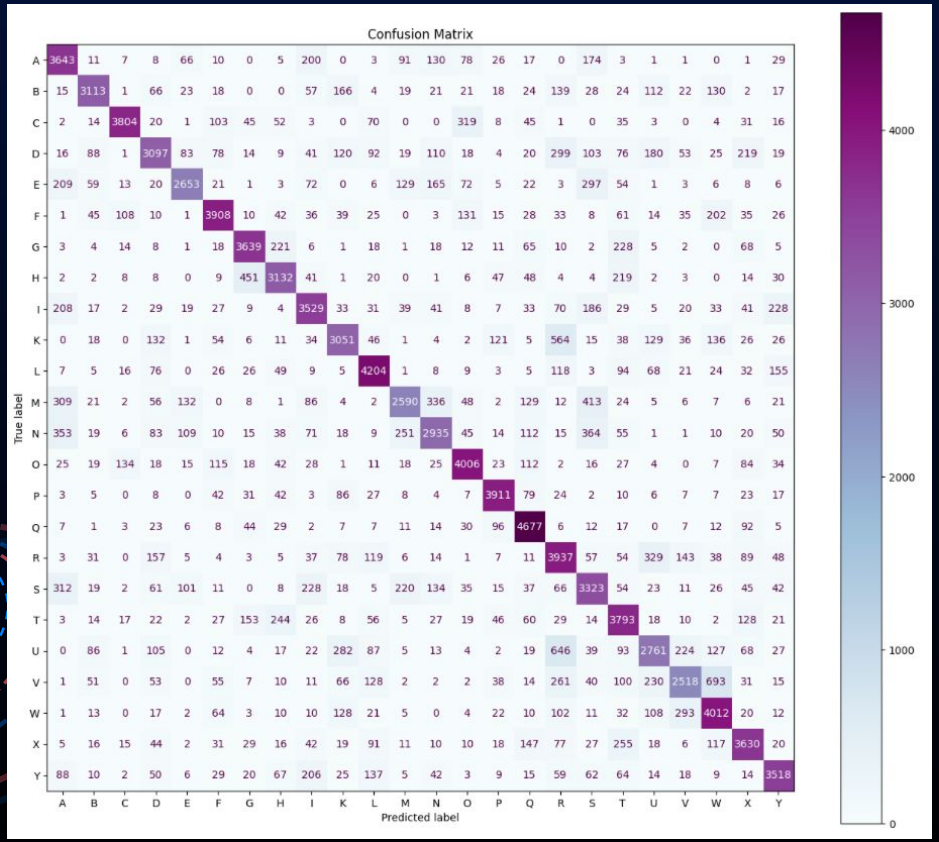
Best Hyperparameters:

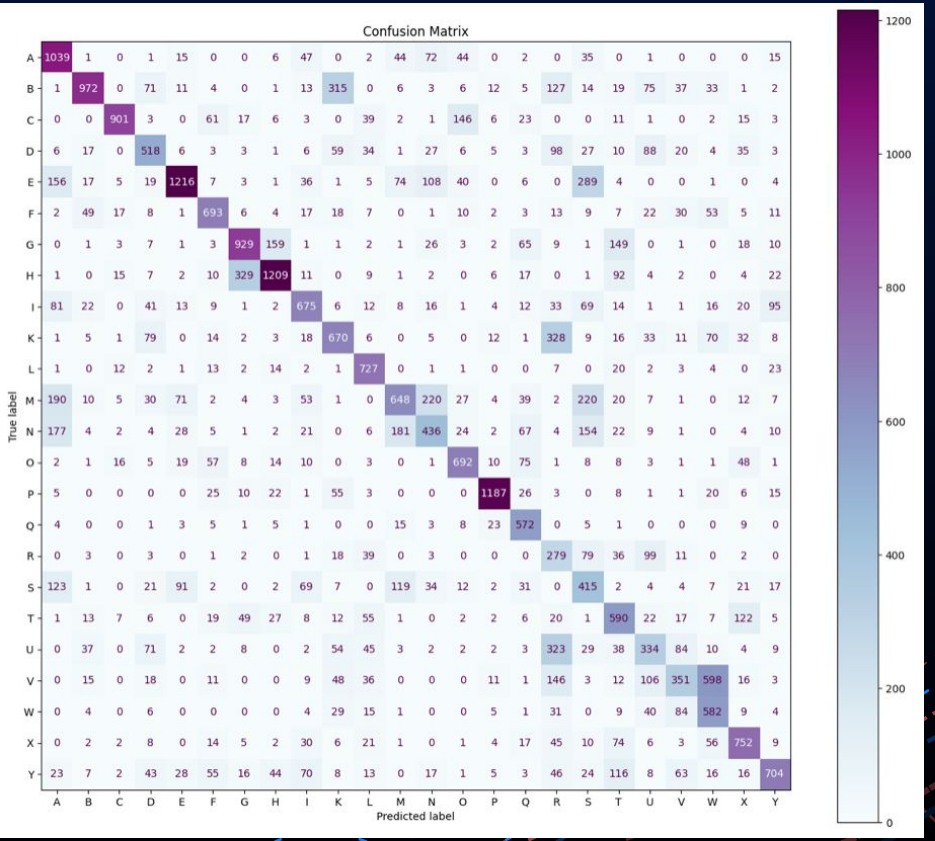
- Subsample : 0.4
- Reg_lamba : 2.25
- Reg_alpha : 2
- Min_child_weight : 30
- Max_depth : 8
- Learning_rate : 0.001
- Gamma : 0
- Colsample_bytree : 0.4

Evaluation Summary

- Accuracy: For train, accuracy is 76% and, and test is 60%
- Precision: A-Y labels range from 0.61-0.92 for train and 0.15-0.91 for test. P & C have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.58-0.85 for train and 0.25-0.87 for test, P & Q have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.7489 for train, indicating a moderate level of agreement between predicted and true labels. For test, it is approximately 0.579.
- Letter R, S, & V had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classes
- Cohen's Kappa is a statistical measure of inter-rater agreement for categorical classifications, which is moderate for this model.

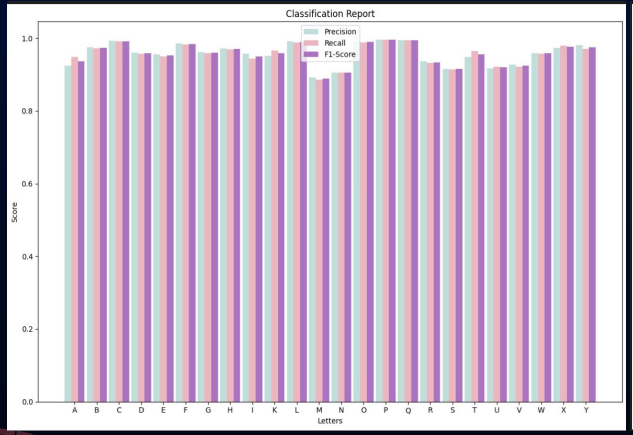
XGBoost: Confusion Matrix

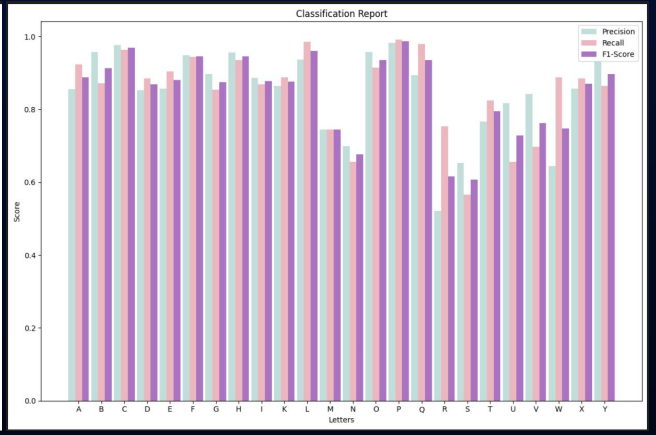




Stacking Ensemble Classifier

Stacking Ensemble: Evaluation Metrics





Estimators:

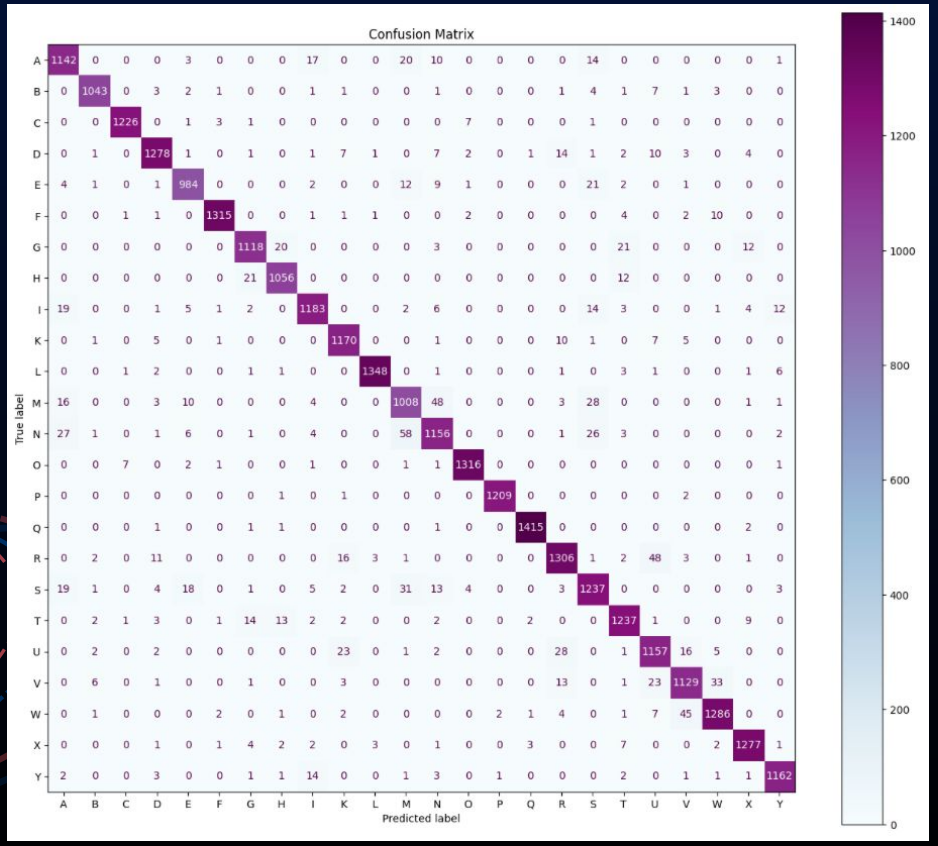
- Logistic Regression
- Support Vector Machine
- Random Forest
- XGBoost

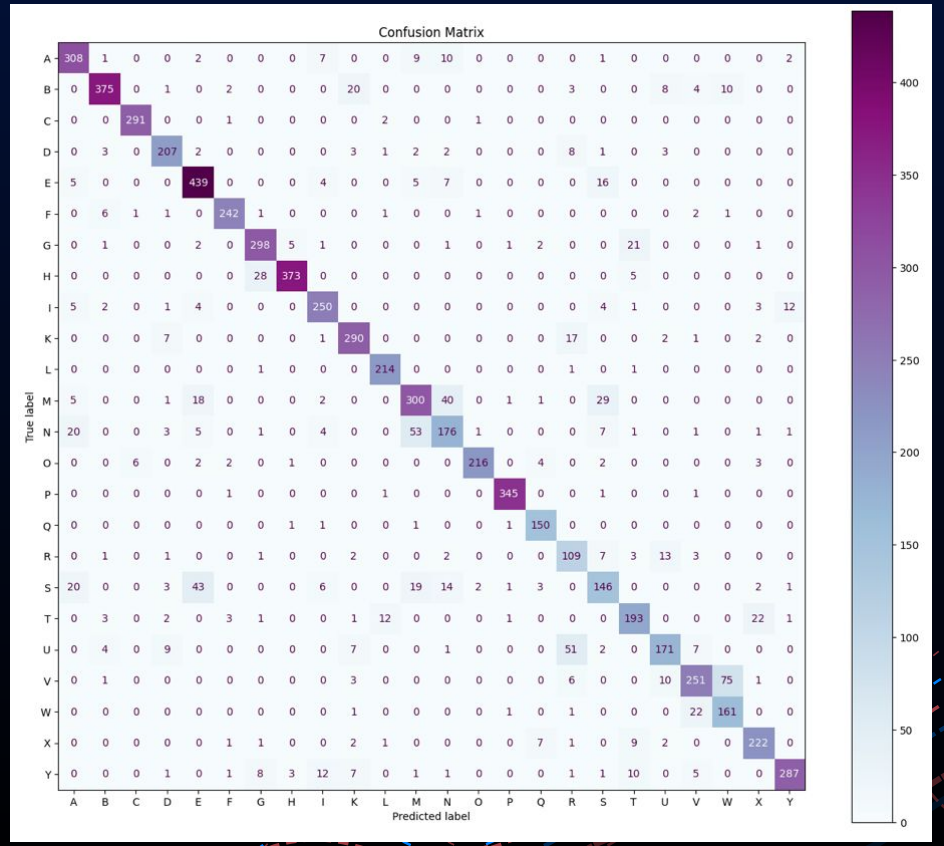
Meta-Estimator: Logistic Regression

Evaluation Summary

- Accuracy: For train, accuracy is 96% and, and test is 86%
- Precision: A-Y labels range from 0.89-1.00 for train and 0.55-0.98 for test. C, O & P have higher precision for test set. Higher values indicates a lower false positive rate
- Recall: The recall ranges from 0.90-1.00for train and 0.64-0.99 for test, P & L have higher recall. Higher values indicates a lower false negative rate
- Support: For test, the values range from 576-1992, indicating varying class frequency
- The Matthews Correlation Coefficient (MCC) is 0.9567 for train, indicating a good higher level of agreement between predicted and true labels. For test, it is approximately 0.852.
- Letter R an S had lower accuracy and showed lower performance in terms of correctly identifying instances of their respective classes
- Cohen's Kappa is a statistical measure of inter-rater agreement for categorical classifications which is high for this model

Stacking Ensemble: Confusion Matrix





Hand Detection

Image Source

Palm/Hand Detection

Image Source

Hand Landmarks

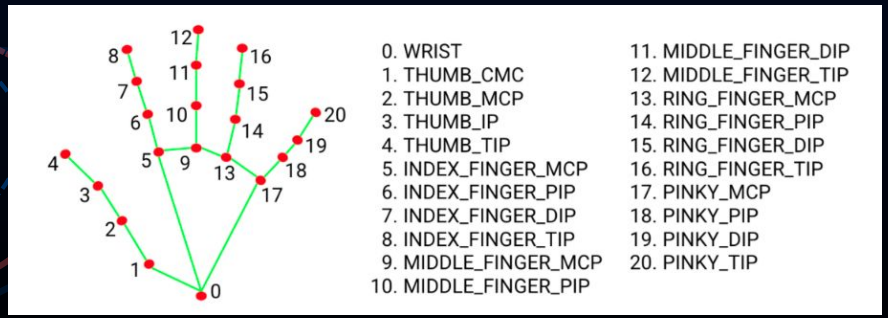


Image Source

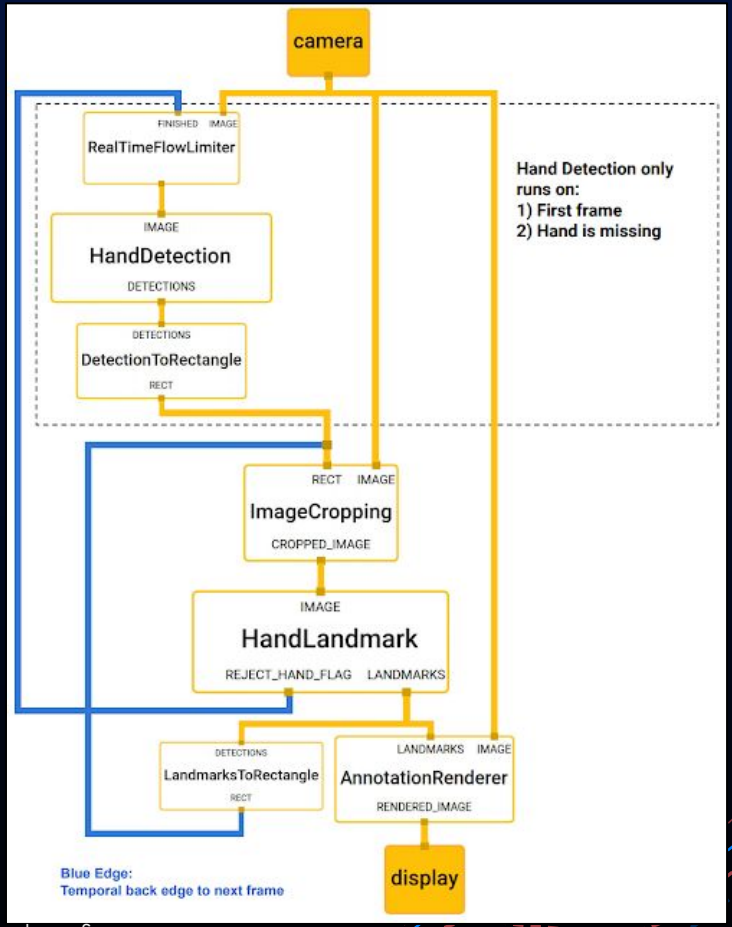


Image Source

Deep Learning

Deep Learning: Approach 1

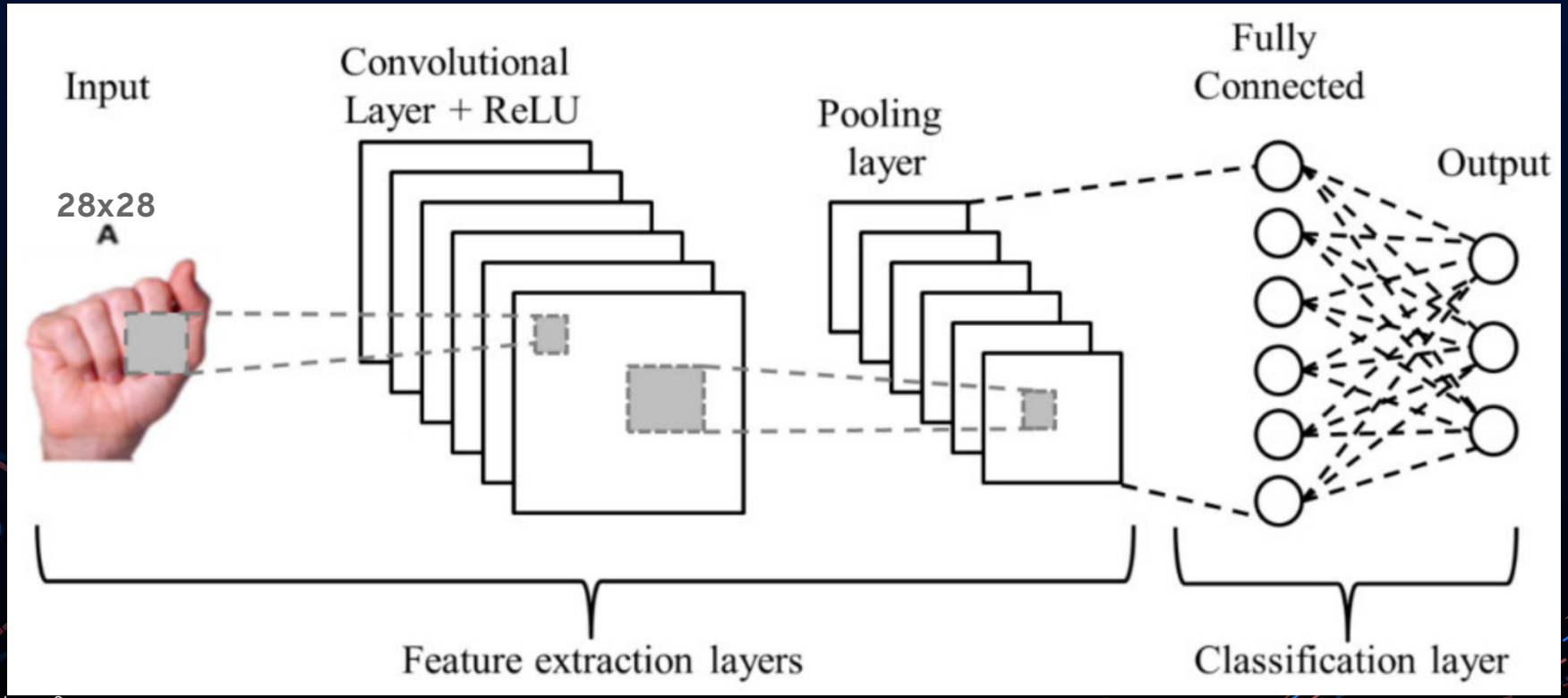
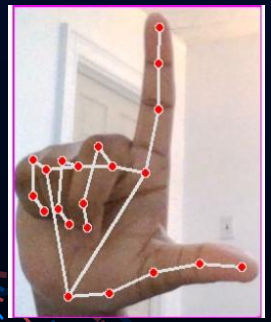


Image Source

Deep Learning: Approach 2

Input Image



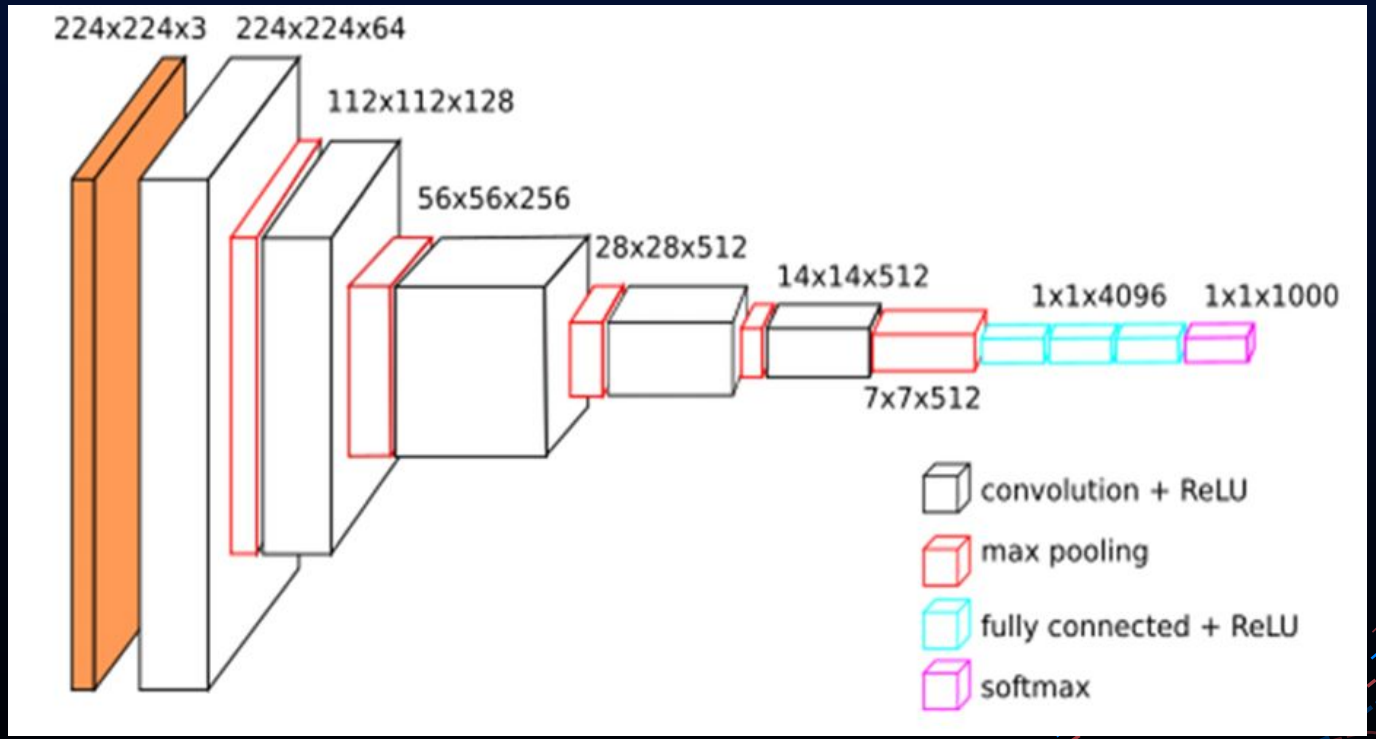


Image Source

Deep Learning: Evaluation Metrics

Accuracy: 0.9879265091863517
Classification report:

	precision	recall	f1-score	support
A	1.00	1.00	1.00	200
B	1.00	1.00	1.00	306
C	1.00	1.00	1.00	201
D	1.00	1.00	1.00	200
E	1.00	1.00	1.00	200
F	1.00	1.00	1.00	200
G	1.00	1.00	1.00	286
H	1.00	1.00	1.00	222
I	1.00	1.00	1.00	206
K	1.00	1.00	1.00	241
L	1.00	1.00	1.00	227
M	1.00	1.00	1.00	215
N	1.00	1.00	1.00	211
O	1.00	1.00	1.00	249
P	1.00	1.00	1.00	262
Q	1.00	1.00	1.00	253
R	1.00	0.75	0.86	258
S	1.00	1.00	1.00	254
T	1.00	1.00	1.00	280
U	0.78	1.00	0.88	231
V	0.99	1.00	0.99	231
W	0.99	0.99	0.99	242
X	1.00	1.00	1.00	274
Y	1.00	1.00	1.00	266
accuracy			0.99	5715
macro avg	0.99	0.99	0.99	5715
weighted avg	0.99	0.99	0.99	5715

MCC: 0.9875267343564018
Cohen's Kappa: 0.9873933134353132

Accuracy: 0.9608355091383812
Classification report:

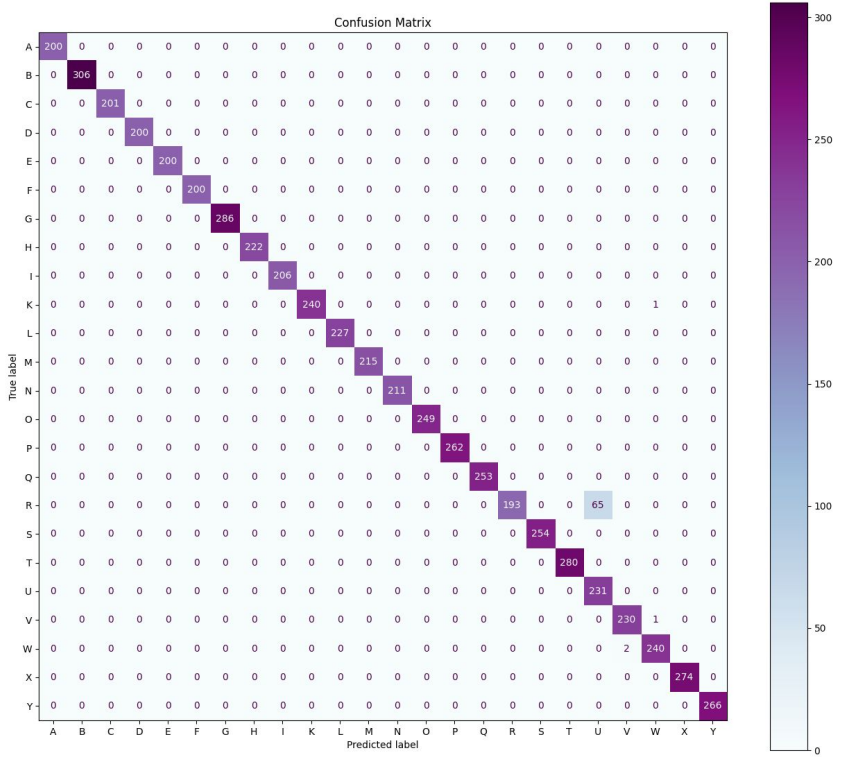
	precision	recall	f1-score	support
A	1.00	0.94	0.97	16
B	1.00	1.00	1.00	16
C	1.00	1.00	1.00	16
D	1.00	1.00	1.00	16
E	1.00	1.00	1.00	16
F	1.00	1.00	1.00	16
G	1.00	1.00	1.00	16
H	1.00	1.00	1.00	16
I	1.00	1.00	1.00	16
K	1.00	1.00	1.00	16
L	1.00	1.00	1.00	16
M	0.94	1.00	0.97	16
N	1.00	1.00	1.00	16
O	1.00	1.00	1.00	16
P	1.00	1.00	1.00	16
Q	1.00	1.00	1.00	16
R	1.00	0.12	0.22	16
S	1.00	1.00	1.00	16
T	1.00	1.00	1.00	15
U	0.53	1.00	0.70	16
V	1.00	1.00	1.00	16
W	1.00	1.00	1.00	16
X	1.00	1.00	1.00	16
Y	1.00	1.00	1.00	16
accuracy			0.96	383
macro avg	0.98	0.96	0.95	383
weighted avg	0.98	0.96	0.95	383

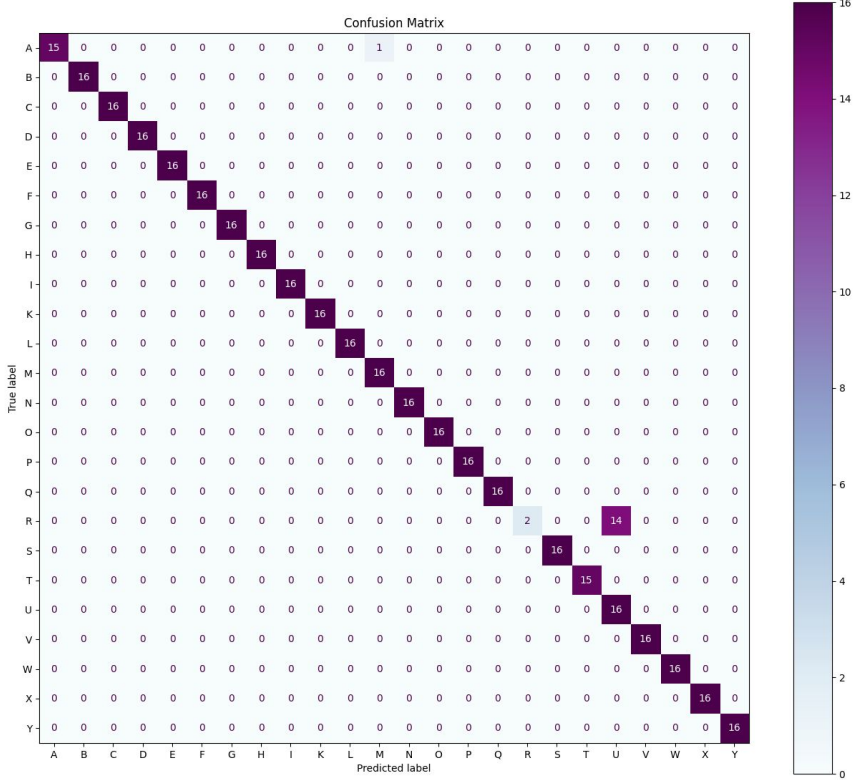
MCC: 0.9604793649086962
Cohen's Kappa: 0.9591324265877532

Evaluation Summary

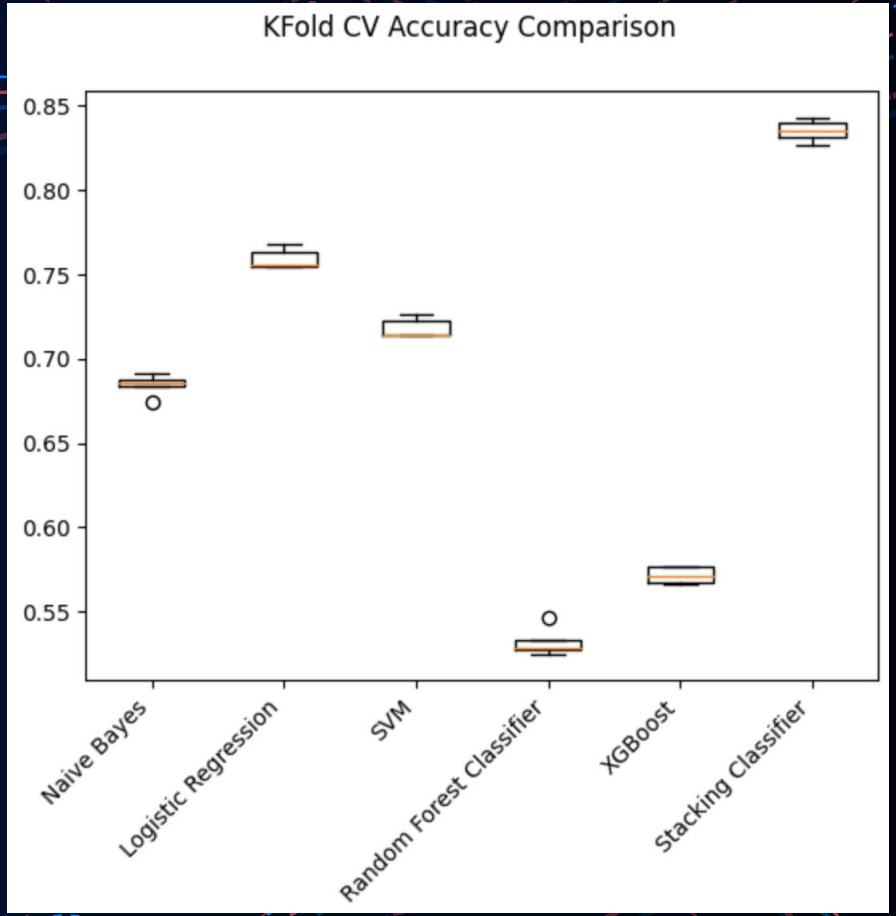
- Accuracy: The accuracy of the model is 0.9879, which means it correctly classified 98.79% of the samples in the dataset.
- Precision: In this case, the precision for all classes except "R," "U," "V," and "W" is 1.00, indicating high precision
- Recall: Like precision, most classes have a recall of 1.00, indicating that the model performed well in capturing positive samples.
- Support: It represents the number of samples in each class.
- F1-score: It is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. Similar to precision and recall, the F1-score for most classes is 1.00, indicating excellent performance.

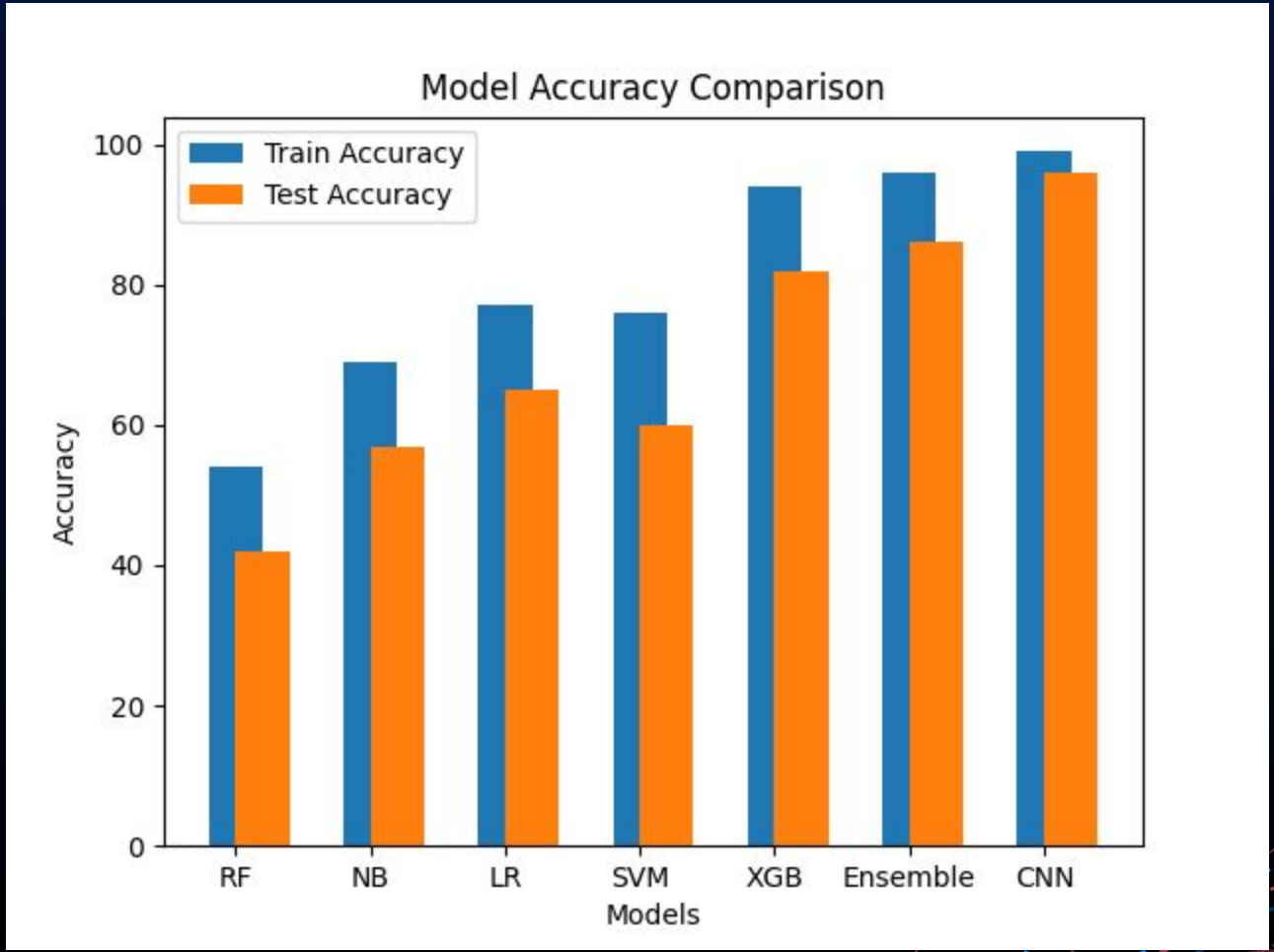
Deep Learning: Confusion Matrix





04. Evaluation


KFold CV Accuracy Comparison
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Naive Bayes
Logistic Regression
SVM
Random Forest Classifier
XGBoost
Stacking Classifier


Model Accuracy Comparison
Train Accuracy
Test Accuracy
Accuracy
100
80
60
40
20
0
RF
NB
LR
SVM
XGB
Ensemble
CNN
Models

05. Demo

06. Conclusions

Best Performing Models for Sign Language Classification

Stacking Ensemble Classifier

Estimators:

- Logistic Regression
- Support Vector Machine
- Random Forest
- XGBoost

Meta-Estimator:

- Logistic Regression

Data:

- 100k Augmented 28x28 Images

CNN Trained on 224 x 244 Images

- CNN consists of a total of 4 hidden layers. These hidden layers are the convolutional layers with 32, 64, 128, and 256 filters respectively
- Each convolutional layer is followed by a max pooling layer to downsample the feature maps.
- The last dense layer consists of 24 units, representing the number of classes in the classification task.
- To overcome with the problem of overfitting, dropout was used to drop the 20% of weights

Conclusions

Data Quality

The performance of models is highly dependent on the quality of data and the ability to extract informative features

Simulated Data ≠ Real World

Training data should be as close as possible to the real world environment.

Certain Models Perform Better with Certain Features

Models have certain decision boundary characteristics, which work better for certain features

Feature Engineering Can Result in More Robust Features

Features that can clarify information like positioning of the fingers will benefit model performance and generality

Thank You