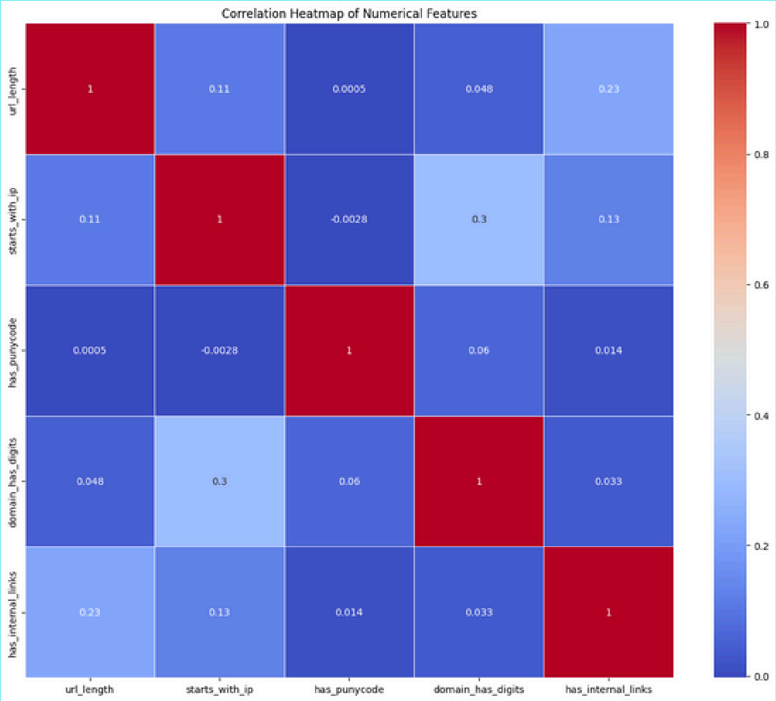
A Comparative Study of ML Approaches for Diverse Data Types Across Multiple Domains

Presented by Group 6

Hemansi Bhalani, Ishana Jabbar, Mehweesh Kadegaonkar, Muhammed Roshan Palayamkot, Shakib Moolur

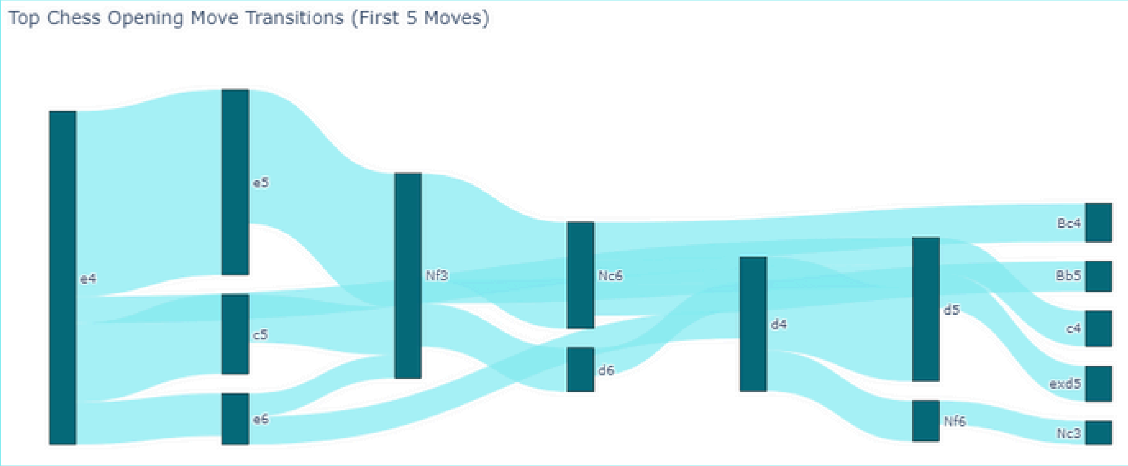
Preprocessing

URL Phishing Dataset



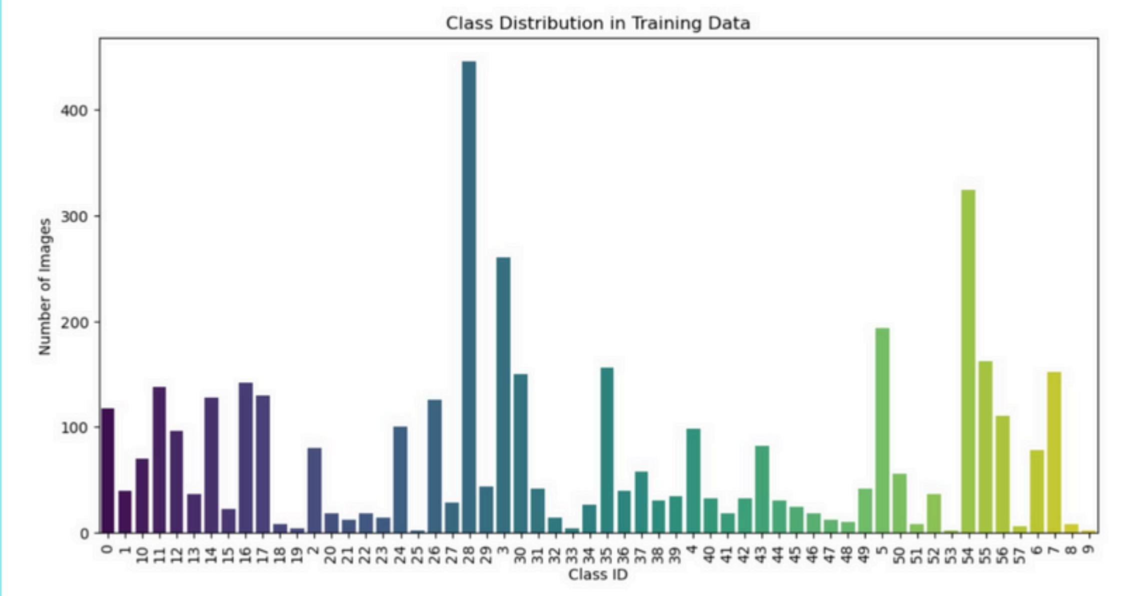
Rows with NULL values were dropped and subsets of balanced data were taken. Preprocessing was done individually for each model if required.

Chess Openings Dataset



The preprocessing is done by cleaning the dataset. This involves removing irrelevant columns, label-encoding categorical features, defining the target variable, and splitting the dataset for training and testing.

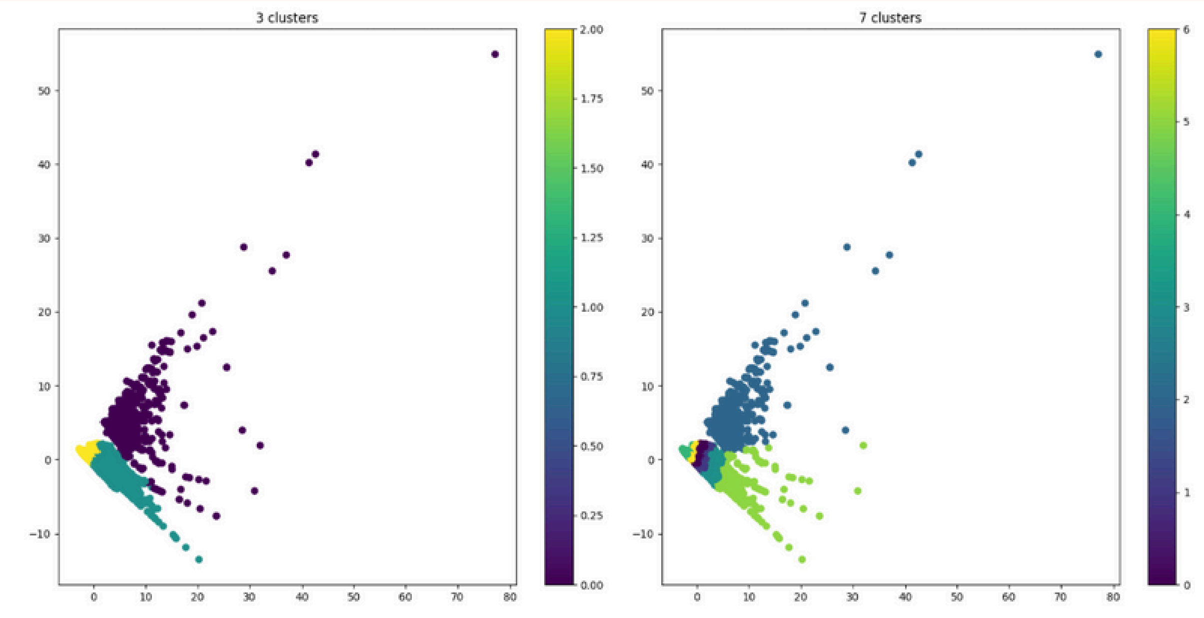
Traffic Sign Dataset



Standard preprocessing such as resizing, converting to grayscale, normalizing, histogram equalization and balancing was performed on the image dataset.

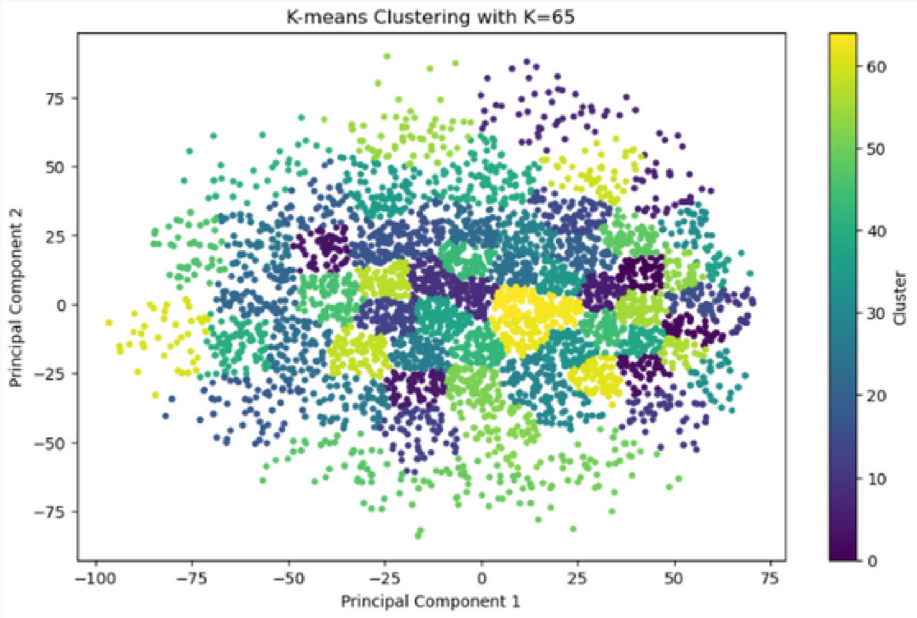
Clustering

The clustering analysis of the datasets was conducted using three primary techniques: K-Means, DBSCAN, and Hierarchical Clustering.



URL Phishing Dataset

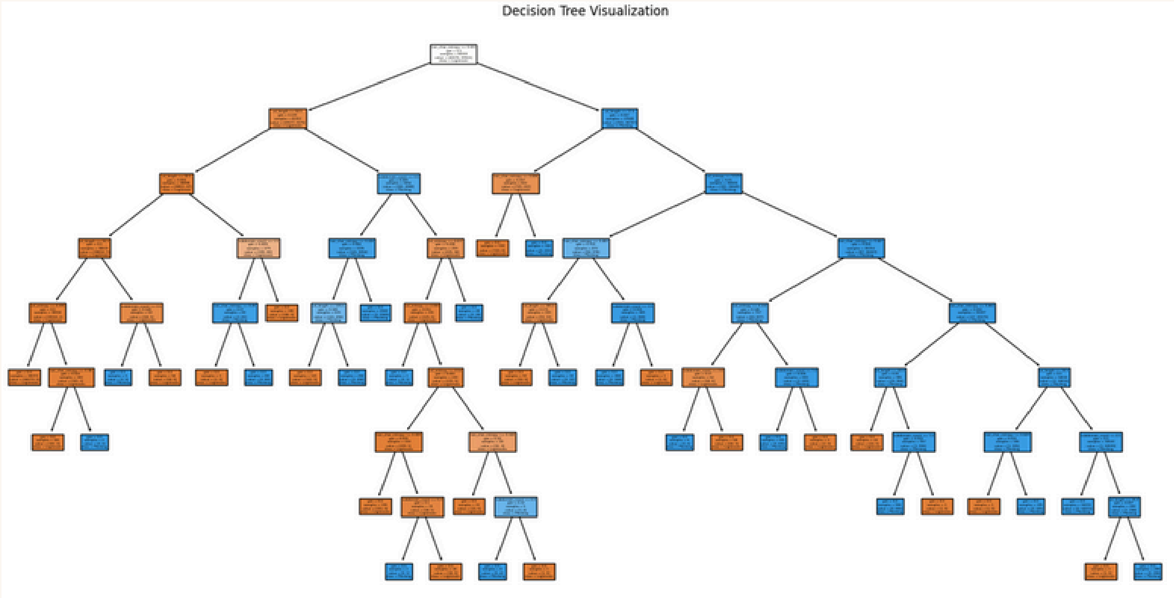
Chess Opening Dataset



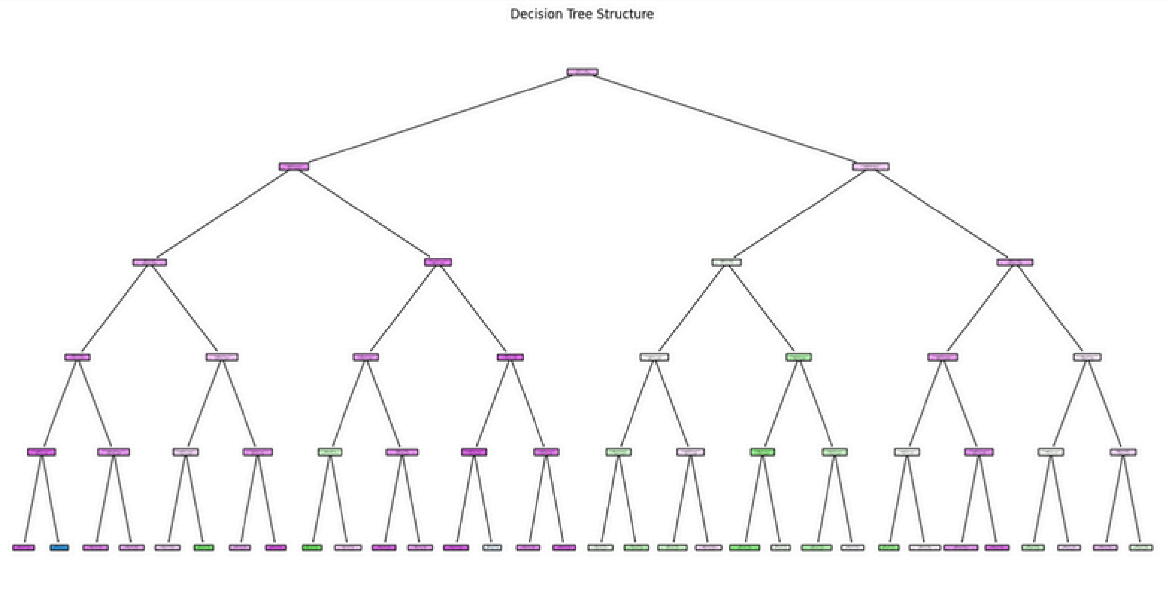
Traffic Sign Dataset

From the results we achieved we observed that K-Means works well with Mixed dataset (Textual, Categorical, Numerical, Boolean) and with sequential dataset i.e. time-series-like data. Whereas DBSCAN works well with recognizing patterns from data and is hence more suitable for an image dataset.

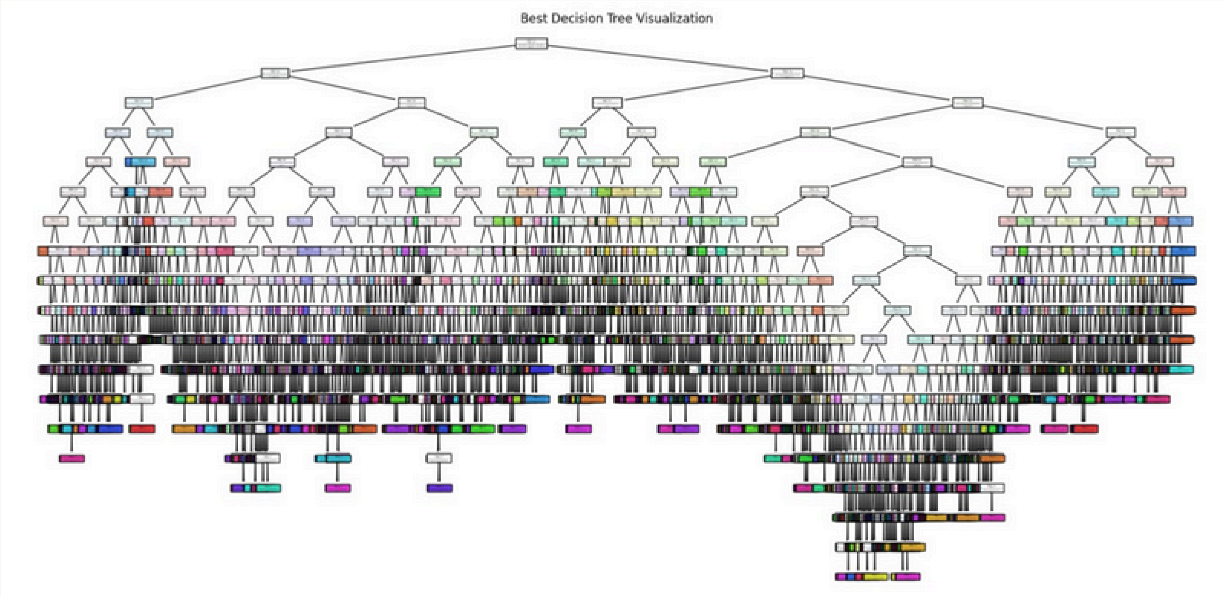
Basic Classifiers and Decision trees



URL Phishing Dataset



Chess Opening Dataset



Traffic Sign Dataset

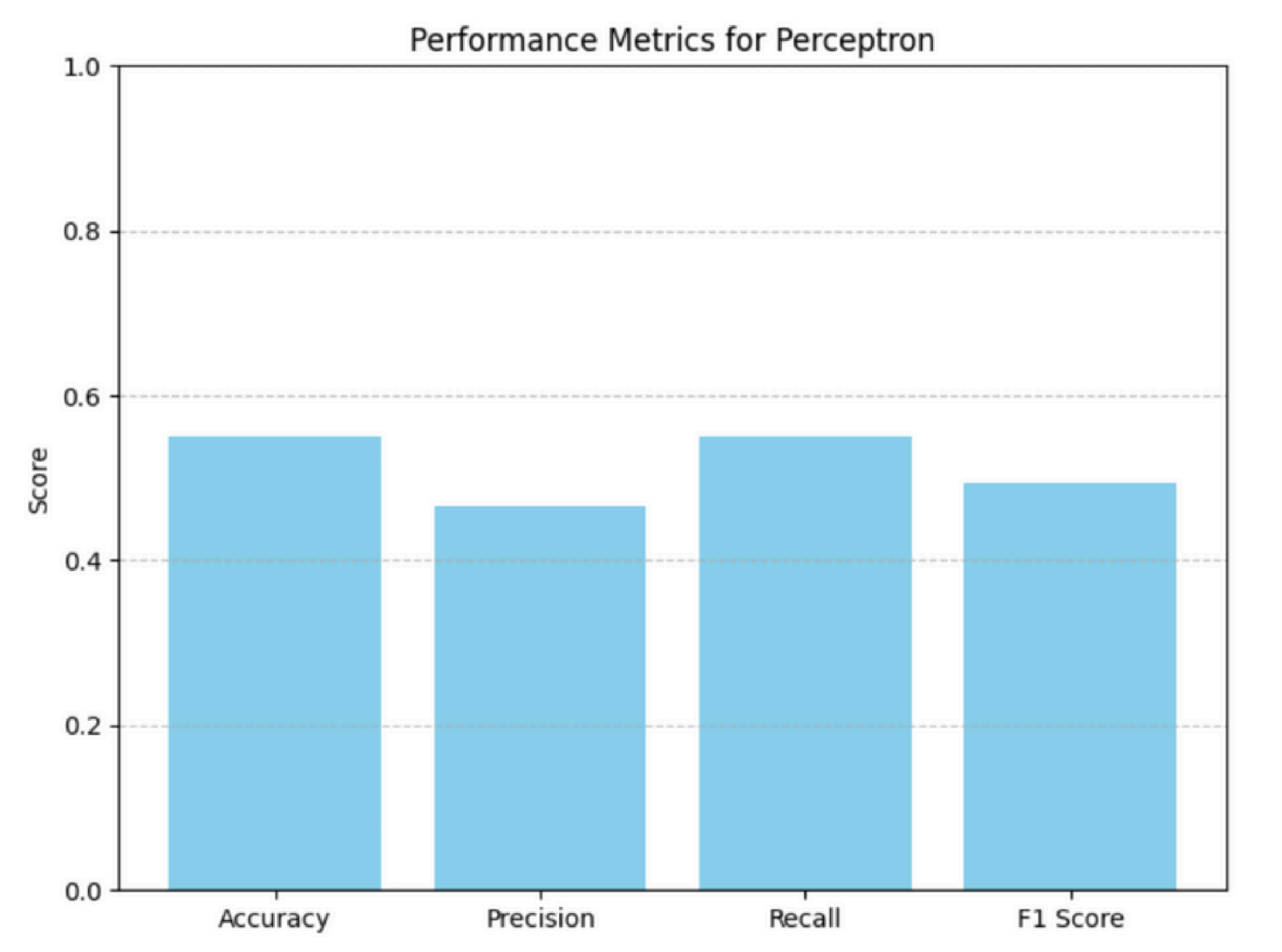
The basic classifiers and decision trees were one of the best performing models for the chosen mixed dataset, while they demonstrated an under average performance for the selected image dataset.

Basic Classifiers and Decision trees

Naive Bayes Algorithm	Accuracy	TP	FP	TN
Gaussian Naive Bayes	0.82265	6809	278	9644
Multinomial Naive Bayes	0.74710	7281	2261	7661
Complement Naive Bayes	0.74705	7281	2262	7660
Bernoulli Naive Bayes	0.70145	7097	2990	6932

Naive Bayes Algorithm	Sensitivity	Specificity	Area Under ROC Curve
Gaussian Naive Bayes	0.675630	0.971981	0.978336
Multinomial Naive Bayes	0.722465	0.772123	0.859816
Complement Naive Bayes	0.722465	0.772022	0.859320
Bernoulli Naive Bayes	0.704207	0.698649	0.780856

Naive Bayes - URL Phishing



Perceptron Metrics - Chess

Neural Networks and CNNs

The Multi-Layer Perceptrons (MLPs) models worked well on our current mixed dataset, demonstrating that MLP is quite strong on structured data, but performed relatively poor when it came to our image dataset as the MLP flattens the image into 1D array, failing to capture spatial relationships which are needed for image recognition. In contrast, Convolutional Neural Networks (CNNs) perform much better on an image dataset because they possess a capability for the extraction of spatial features in a more effective way.

```
Epoch 60/60
150/150 ━━━━━━━━━━━━━━━━━━━━ 2s 14ms/step - accuracy: 0.9612 - loss: 0.1209 - val
_accuracy: 0.8975 - val_loss: 0.4052
38/38 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step
Current model metrics: {'accuracy': 0.8975, 'precision': 0.9041650091752785, 'rec
all': 0.8975, 'f1_score': 0.8973543792422718}
Current model is better. Replacing the last saved model.
```

Traffic Sign Dataset - CNN

Test Accuracy with Noisy Labels and Features: 0.9491

	precision	recall	f1-score	support
0	0.95	0.95	0.95	9995
1	0.95	0.95	0.95	10005
accuracy			0.95	20000
macro avg	0.95	0.95	0.95	20000
weighted avg	0.95	0.95	0.95	20000

URL Phishing - MLP

```
Performance Metrics for All Architectures:
Model 1 - Architecture: {'layers': 2, 'units': 512, 'dropout_rate': 0.5}
  Accuracy: 0.2700
  Precision: 0.3327
  Recall: 0.2700
  F1 Score: 0.2846
  Sensitivity: 0.2700
Model 2 - Architecture: {'layers': 3, 'units': 512, 'dropout_rate': 0.4}
  Accuracy: 0.2595
  Precision: 0.3311
  Recall: 0.2595
  F1 Score: 0.2730
  Sensitivity: 0.2595
Model 3 - Architecture: {'layers': 4, 'units': 512, 'dropout_rate': 0.4}
  Accuracy: 0.2015
  Precision: 0.2226
  Recall: 0.2015
  F1 Score: 0.1799
  Sensitivity: 0.2015
```

Traffic Sign Dataset - MLP

Presented by Group 6

Thank you very much!