Friday 29/11/2024

Machine learning techniques to detect lung cancer

Data Mining and Machine Learning - F21DL

UG - 11

Authors:
Yasmeen Jasim H00387810
Lourde Hajjar H00376828
Abdallah Moosa H00416466
Lukas Kras H00390699
Ashar Ejaz H00367838

PROJECT OVERVIEW

Project Goal: Detect Lung cancer using machine learning techniques

Motivation: Lung cancer is a leading cause of cancer related deaths; early detection improves survival rates.

Dataset 1: Binary tabular, 16 attributes, 284 instances.
Dataset 2: categorical tabular, balanced distribution 26 attributes with 1000 instances.
Dataset 3: image dataset with 1190 images (benign, malignant, normal classes)

Machine learning techniques:

- Data analysis and preprocessing
- clustering
- classification
- Neural networks

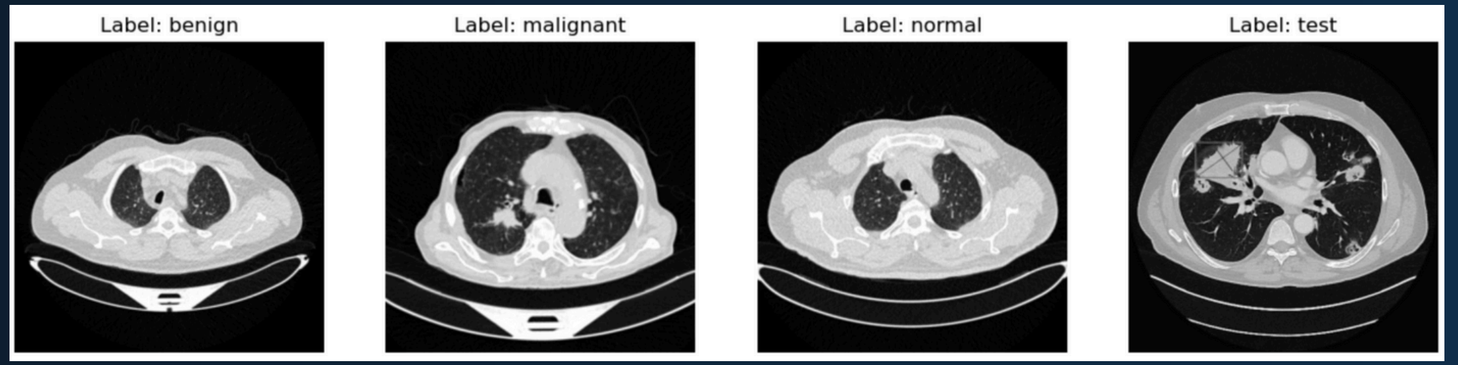
Binary Dataset Sample:

	GENDER	AGE	SMOKING	YELLOW_FINGERS	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	2	YES
1	M	74	2	1	2	2	YES
2	F	59	1	1	1	2	NO
3	M	63	2	2	2	2	NO
4	F	63	1	2	1	1	NO

Categorical Dataset Sample:

Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Dry Cough	Snoring	Level
33	1	2	4	5	3	4	Low
17	1	3	1	5	7	2	Medium
35	1	4	5	6	7	2	High
37	1	7	7	7	7	5	High
46	1	6	8	7	2	3	High

Image Dataset Sample:



197 images *561 images* *416 images* *120 images*

Dataset Analysis and Preprocessing

Key findings:

- Dataset 1: strongly imbalanced target class, weak feature correlation, somewhat normal age distribution in age range 40-80
- Dataset 2: abnormalities in age range, strong feature correlation to target class, mostly balanced (e.g., Obesity: 0.83)
- Dataset 3: pixel intensity range abnormality, images of different sizes, inconsistent zoom level, normalization through computer vision and resizing, 8% of image get corrupted on edge detection

Dataset variations:

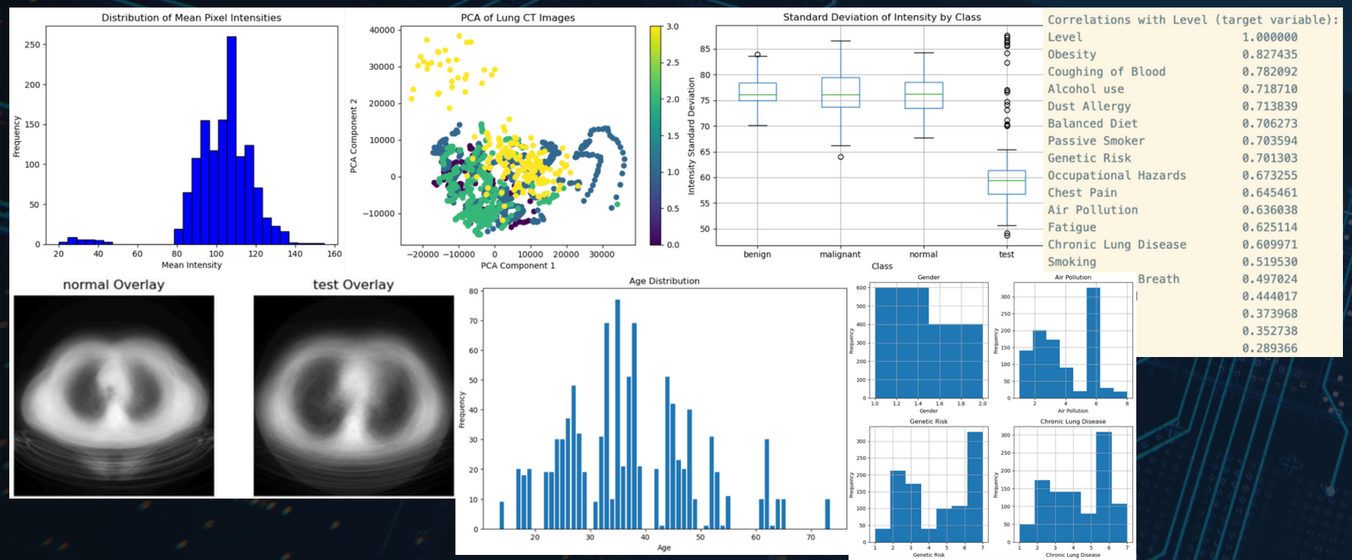
- Balanced and unbalanced versions
- Feature extraction for correlated attributes
- Image normalization and noise reduction
- Image edge detection

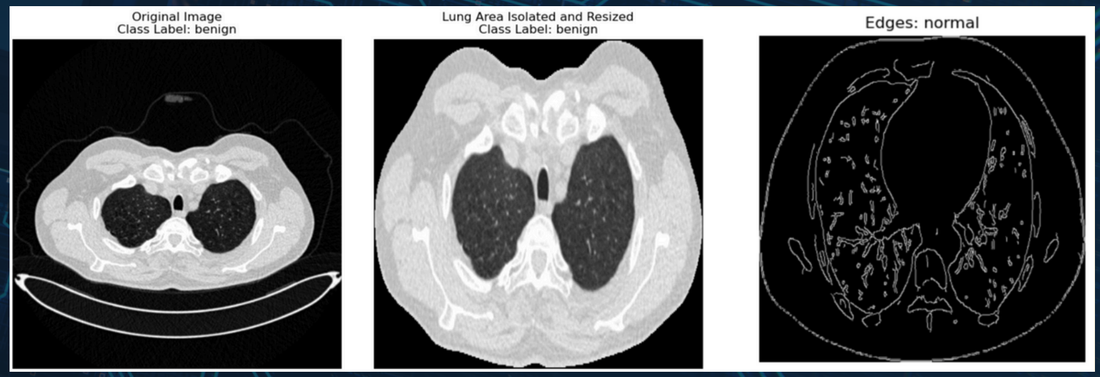
	GENDER
0	1
1	1
2	2
3	1
4	2

AGE	SMOKING
0.727273	0.0
0.803030	1.0
0.575758	0.0
0.636364	1.0
0.636364	0.0

ALLERGY_ALCOHOL CONSUMING_COUGHING	LUNG_CANCER
0.666667	1.0
0.333333	1.0
0.333333	0.0
0.333333	0.0
0.333333	0.0

Columns from different variations of dataset 1

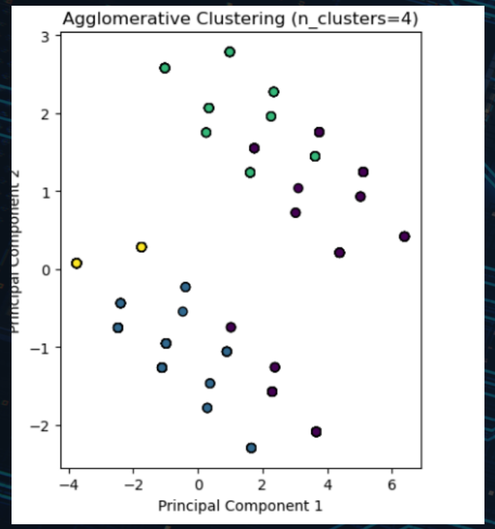




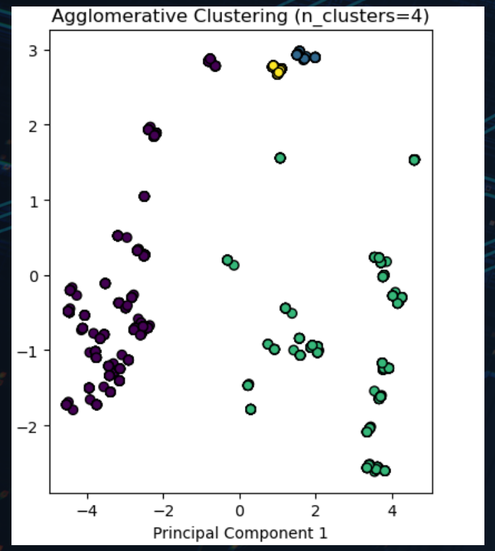
Clustering Techniques

Agglomerative hierarchical clustering

Dataset 1

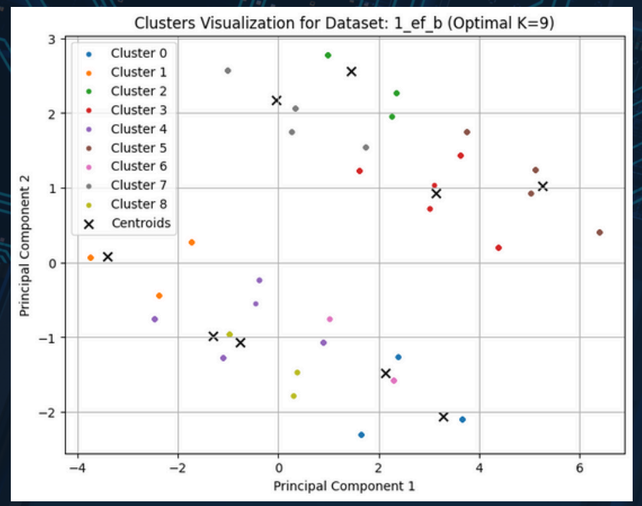


Dataset 2

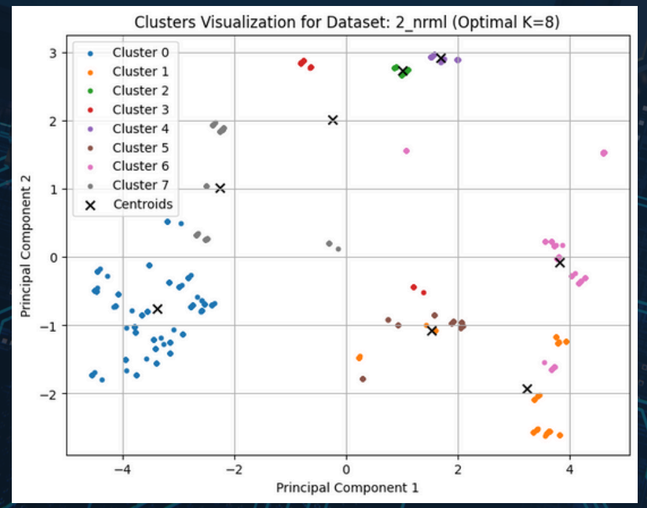


K-Means Clustering

Dataset 1

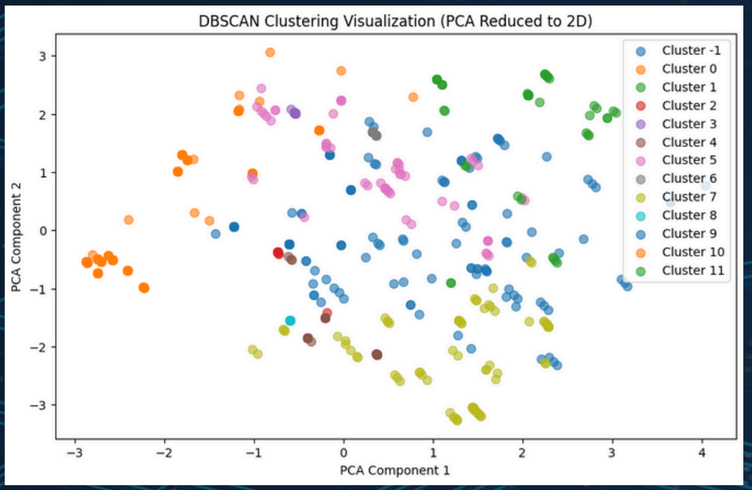


Dataset 2

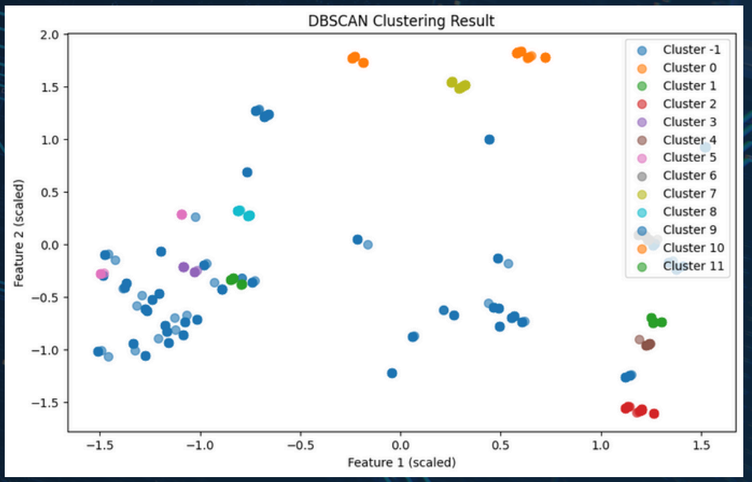


DBSCAN

Dataset 1

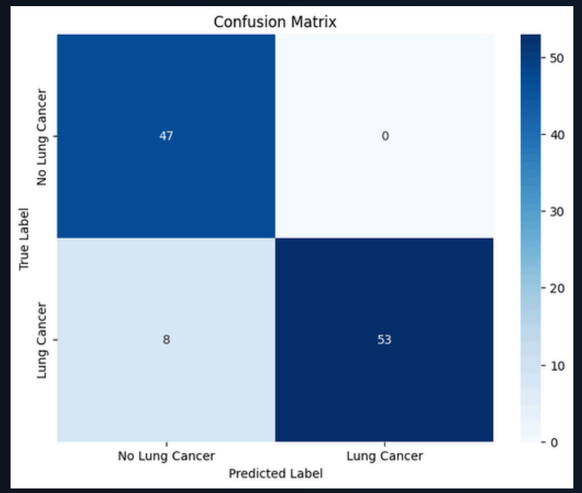


Dataset 2

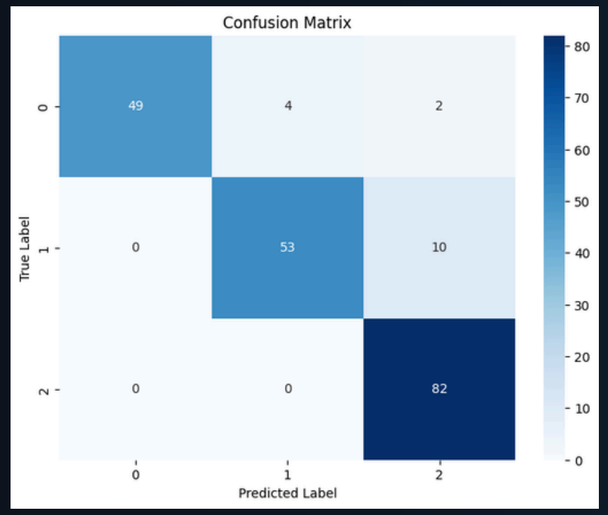


Classification Models

Logistic Regression:

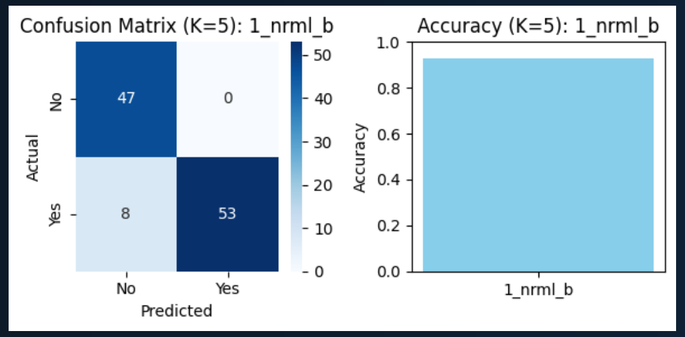


Dataset 1

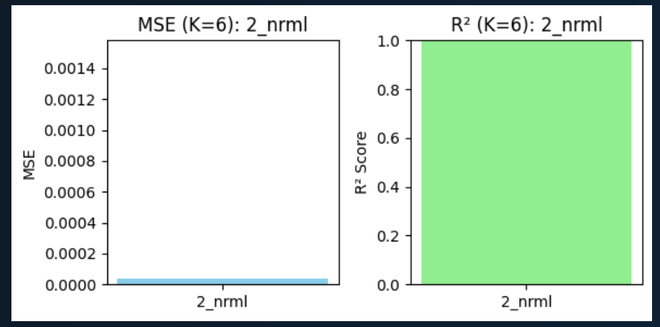


Dataset 2

K-Nearest Neighbor (KNN)



Dataset 1

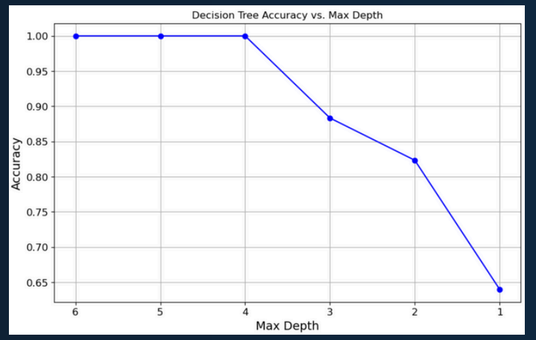


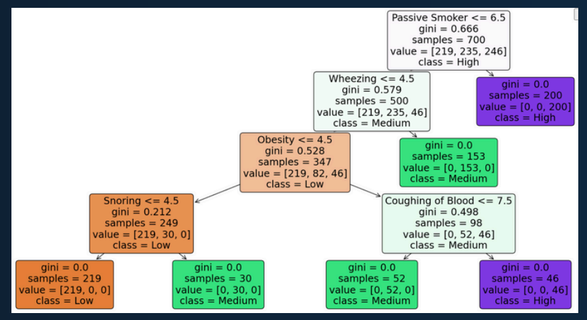
Dataset 2

Decision Trees:

```
Mean Accuracy: 0.9203703703703704
Standard Deviation of Accuracy: 0.02385944208652801

Classification Report:
              precision    recall  f1-score   support

           1       0.86      0.97      0.92        72
           2       0.98      0.88      0.92        90

    accuracy                           0.92       162
   macro avg       0.92      0.93      0.92       162
weighted avg       0.93      0.92      0.92       162
```

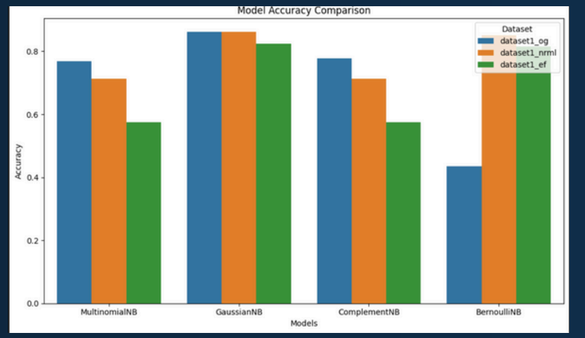
Dataset 1



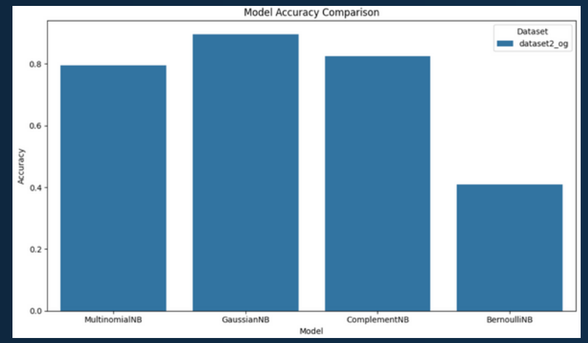


Dataset 2

Naive Bayes



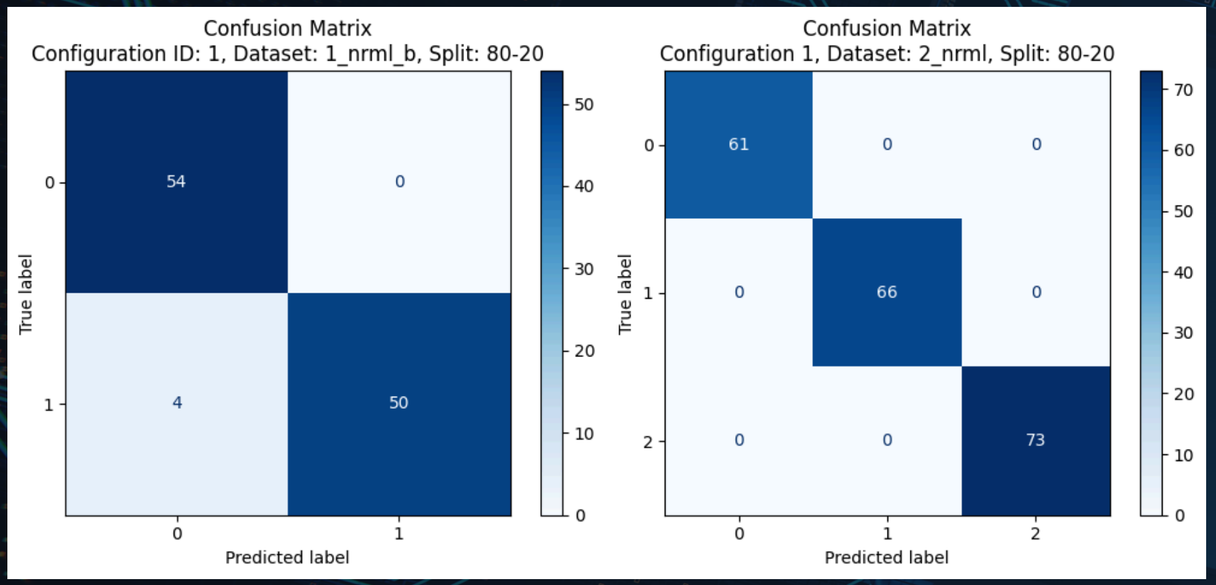
Dataset 1



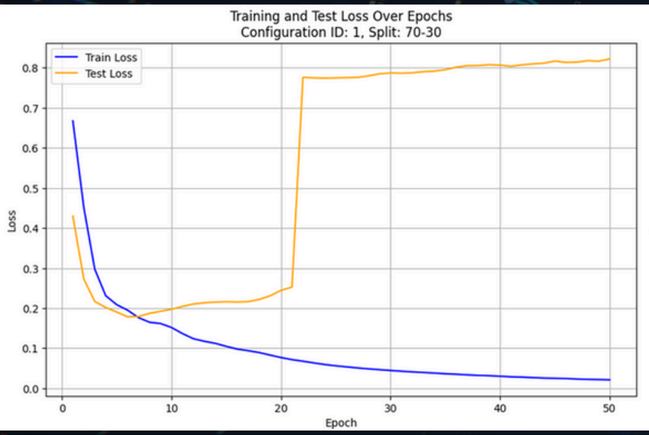
Dataset 2

Neural Networks

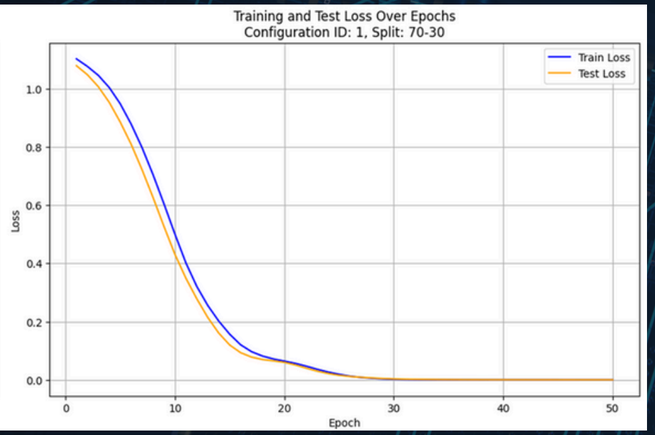
Multi Layer Perceptron (MLP)



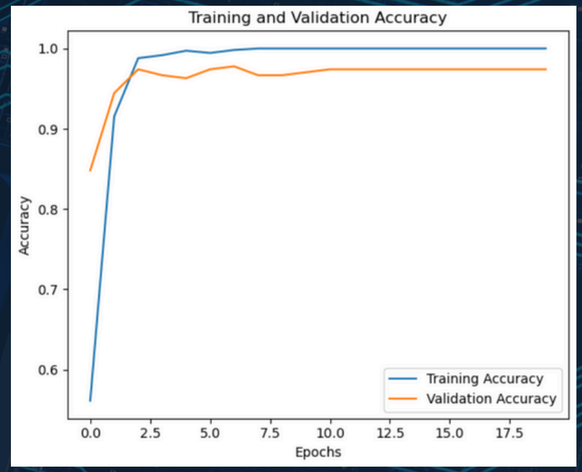
Dataset 1



Dataset 2



Convolutional Neural Network (CNN)



Dataset 3:
Focused variation overfitting on basic cnn

Results Summary:

	Activation Function	Accuracy	Precision	Recall	F1-Score
0	relu	0.792285	0.802644	0.792285	0.788643
1	sigmoid	0.697329	0.729109	0.697329	0.695978
2	tanh	0.649852	0.671807	0.649852	0.604441
3	elu	0.738872	0.767737	0.738872	0.732296

Activation functions testing on the focused variation