Cancer classification based on gene expression data

Our **Agenda** for Today

INDEX

- INTRODUCTION
- MOTIVATION
- DATASET
- DATA PREPROCESSING
- MODEL ARCHITECTURE
- RESULTS
- GRAPHS AND TABLES
- CONCLUSION
- REFERENCES

INTRODUCTION

Cancer classification using gene expression data is a common application of machine learning in bioinformatics. Gene expression data measures the activity level of genes in a sample, and can provide insights into the biological processes occurring in a cell.

To classify cancer using gene expression data, the first step is to preprocess the data to remove noise and normalize the expression levels.

Then, machine learning algorithms such as support vector machines (SVMs), random forests, or neural networks can be trained on the preprocessed data to learn a model that can accurately classify new samples.

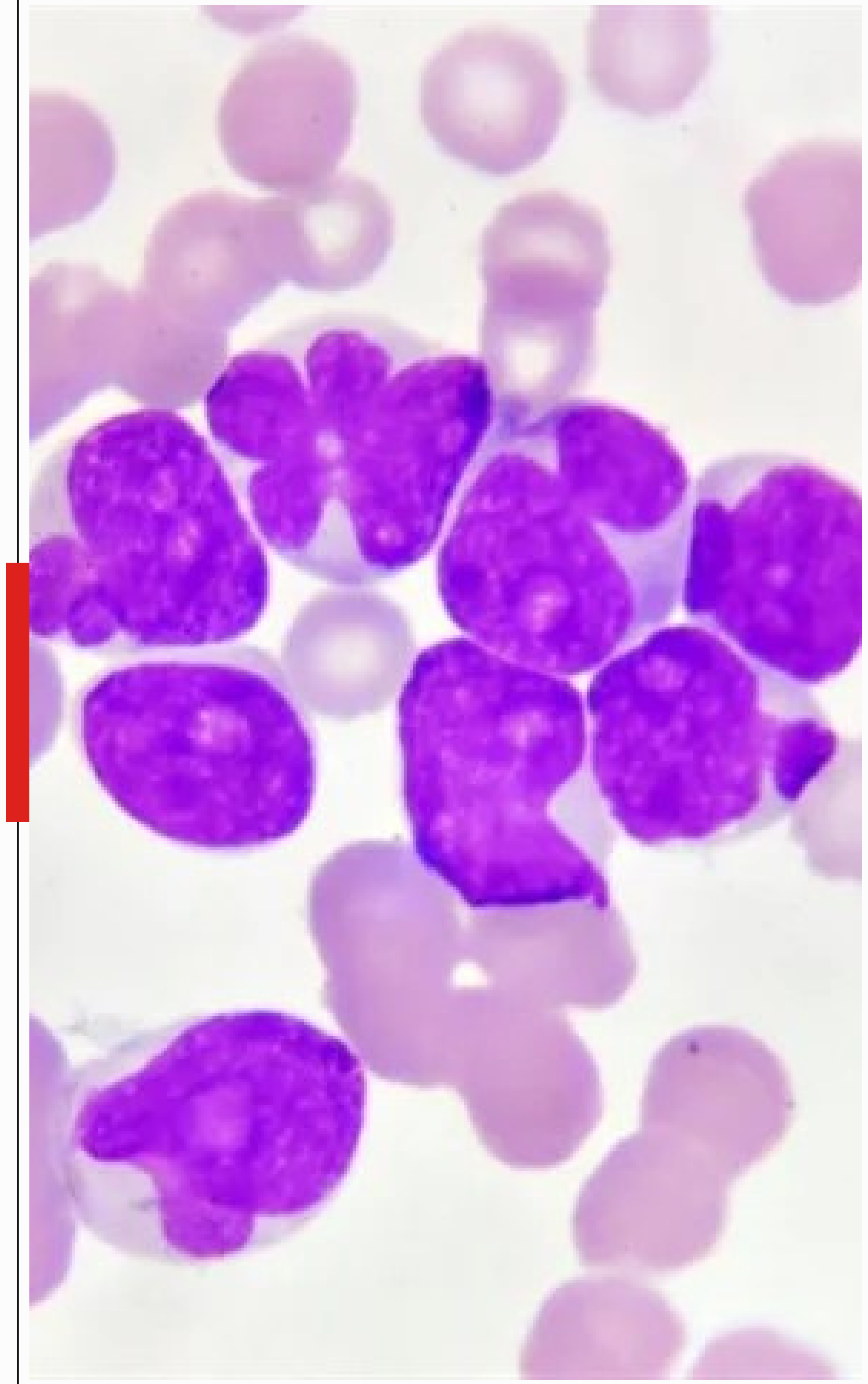
Techniques such as principal component analysis (PCA) can be used to reduce the dimensionality of the data and identify the most important features.

MOTIVATION

Cancer is a life-threatening disease that affects millions of people worldwide, and early detection is critical for successful treatment and survival. The earlier cancer is detected, the more likely it is to be treatable and cured.

Cancer is a complex disease that can develop in different organs and tissues, and can have many different subtypes with different characteristics and treatment options.

The motivation for cancer detection problem is to develop accurate and reliable machine learning models that can assist healthcare professionals in detecting cancer at an early stage and improve patient outcomes. Such models have the potential to save lives, reduce healthcare costs, and advance our understanding of cancer biology and treatment.

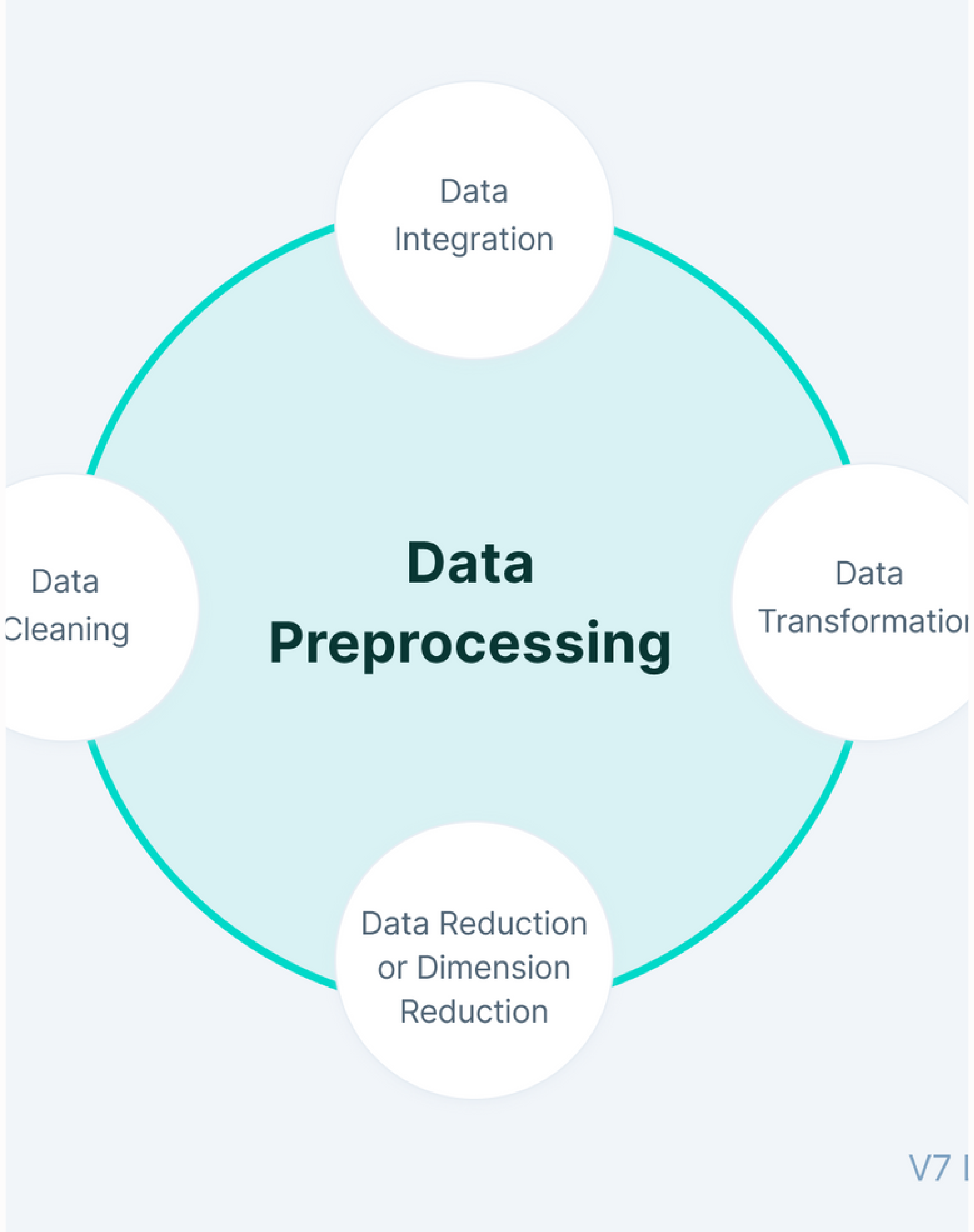


DATASET

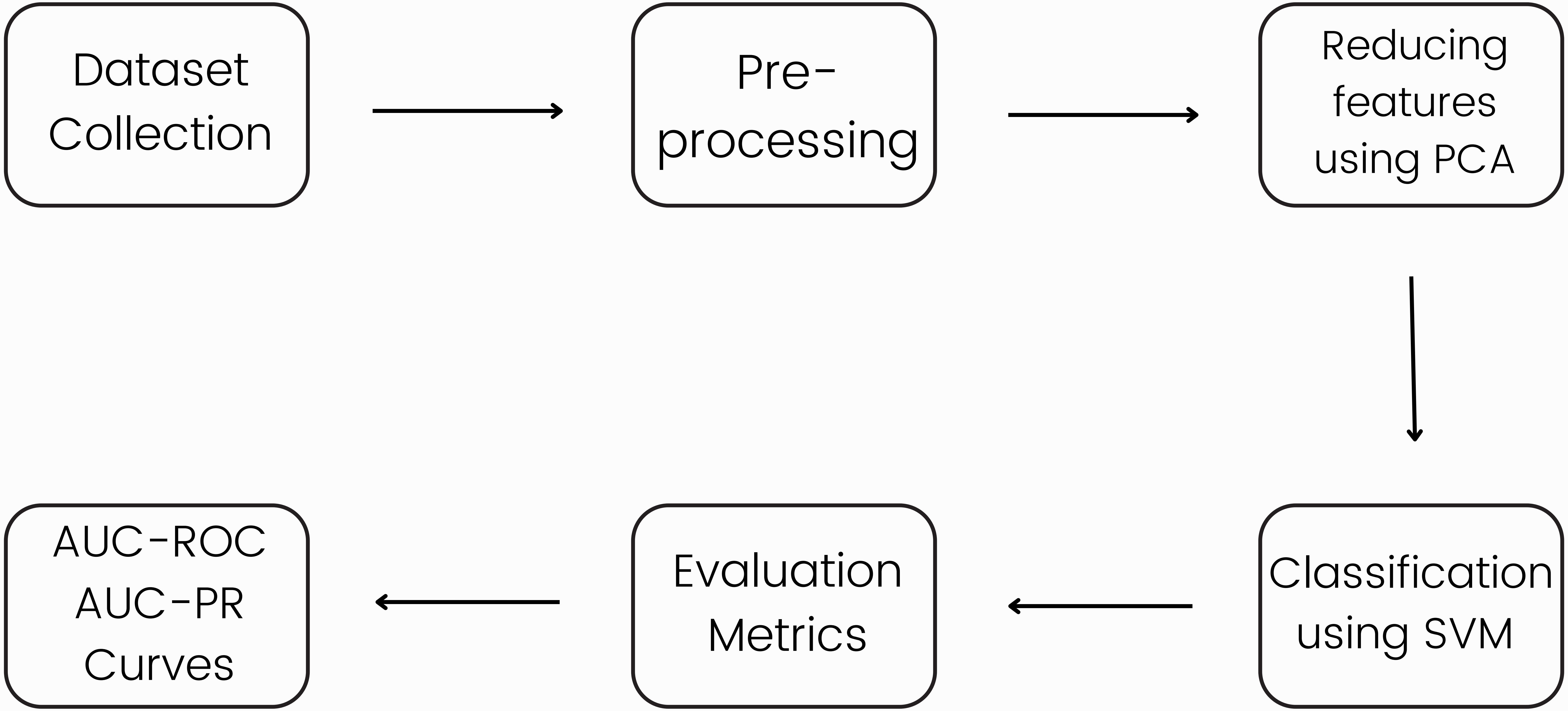
- The dataset has been obtained from Kaggle.
- There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples).
- These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood.
- ALL is a cancer that starts in the lymphoid cells of the bone marrow and affects the production of lymphocytes, a type of white blood cell that helps fight infections. ALL is more common in children than adults and is one of the most common types of childhood cancer.
- AML, on the other hand, is a cancer that starts in the myeloid cells of the bone marrow and affects the production of red blood cells, white blood cells, and platelets. AML is more common in adults than children and is more likely to occur in people over the age of 60.

DATA PREPROCESSING

- Converting class labels into numeric form.
- Removed un-necessary columns.
- Neither the training and testing column names are not in numeric order, so we have re-ordered them into proper order.
- we then transposed the columns and rows so that genes become features and each patient's observations occupies a single row.
- Removed duplicate meaningless row and converted it into a single row.
- Associated the target labels with the right patients.
- created a scaled version of the dataset.
- Reduced the dimensionality from 7129 features to 22 features with 90% variance.
- We picked top 3 PCA components and plotted them.



MODEL ARCHITECTURE



Dataset Collection

Pre-processing

Reducing features using PCA

Classification using SVM

Evaluation Metrics

AUC-ROC AUC-PR Curves

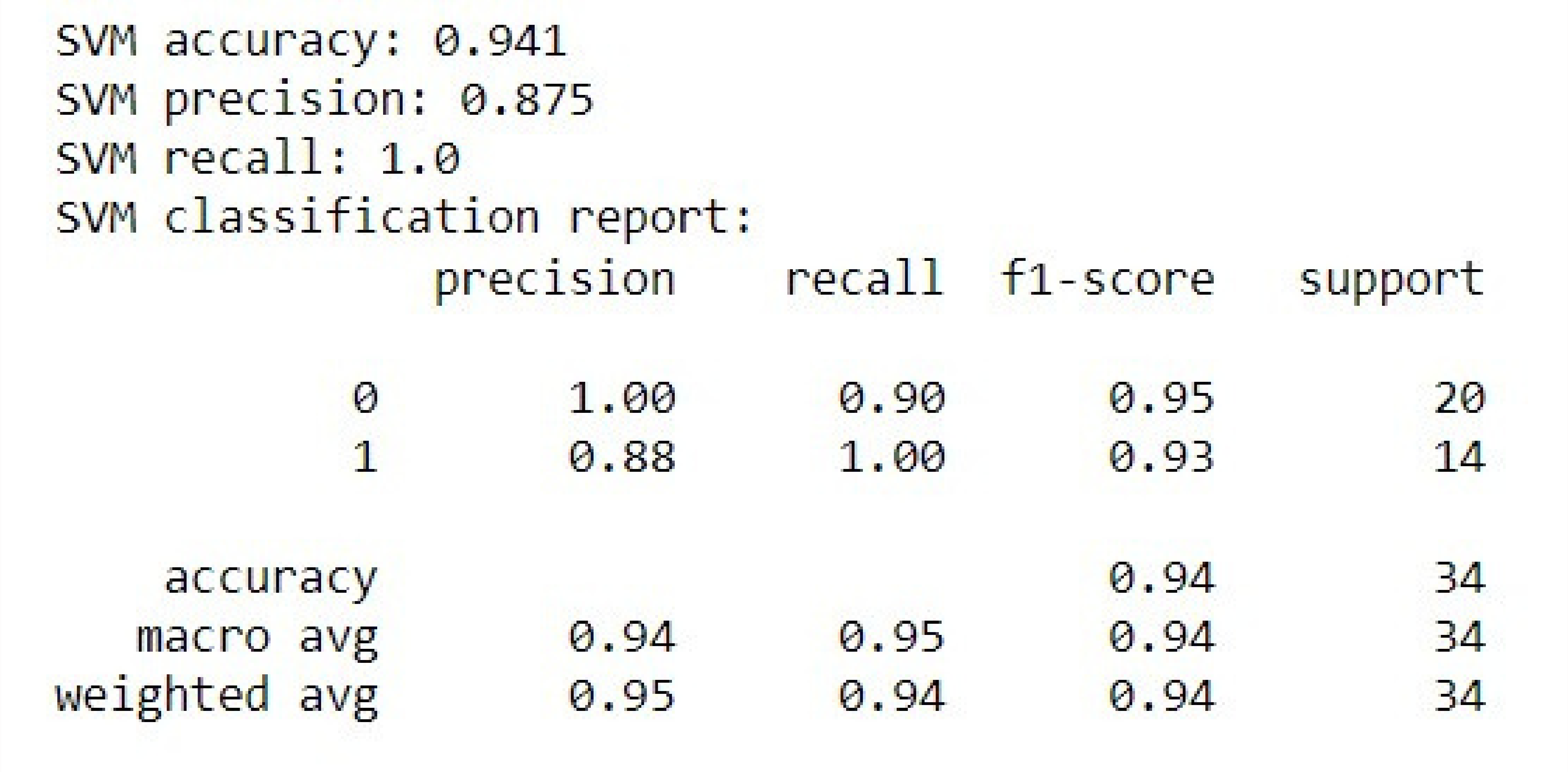
MODEL ARCHITECTURE

Description :

- We have removed all the noise and unwanted rows/columns from the dataset. It is clearly mentioned in the pre-processing slide.
- Later we have used PCA to reduce the features and then picked the first three eigen vectors.
- The classifiers we have used is Support Vector Machine(SVM).
- SVM aims to find the optimal hyperplane that separates the data into different classes or predicts the target variable, by maximizing the margin between the classes and minimizing the classification error.
- Then we found the Evaluation Metrics such as Accuracy, Precision, F1 Score and Recall. We also plotted the confusion matrix of our classifier in order to get the TP, TN, FP, FN values.
- Using the evaluation metrics, we finally plotted the ROC and the precision-recall(PR) curves through which we are able to compare the results.

RESULTS

- While simply predicting everything as acute lymphoblastic leukemia (ALL) we got an accuracy of 0.588.
- But after using the SVM classifier, the results are as follows :
- A good accuracy of 0.941 and a precision of 0.875 is obtained.

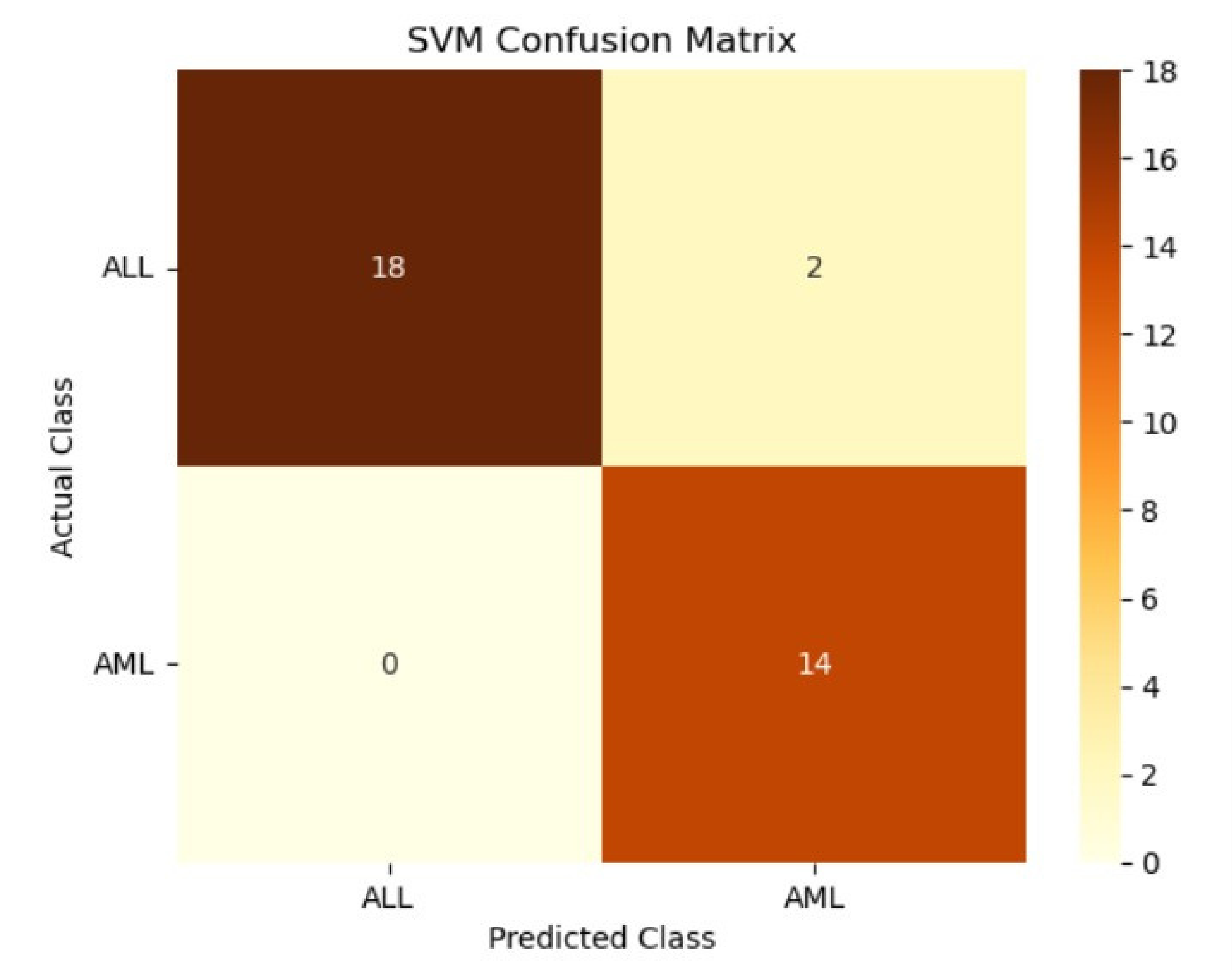


```
SVM accuracy: 0.941
SVM precision: 0.875
SVM recall: 1.0
SVM classification report:
              precision    recall  f1-score   support

           0       1.00      0.90      0.95        20
           1       0.88      1.00      0.93        14

    accuracy                           0.94        34
   macro avg       0.94      0.95      0.94        34
weighted avg       0.95      0.94      0.94        34
```

GRAPHS AND TABLES

CONFUSION MATRIX :

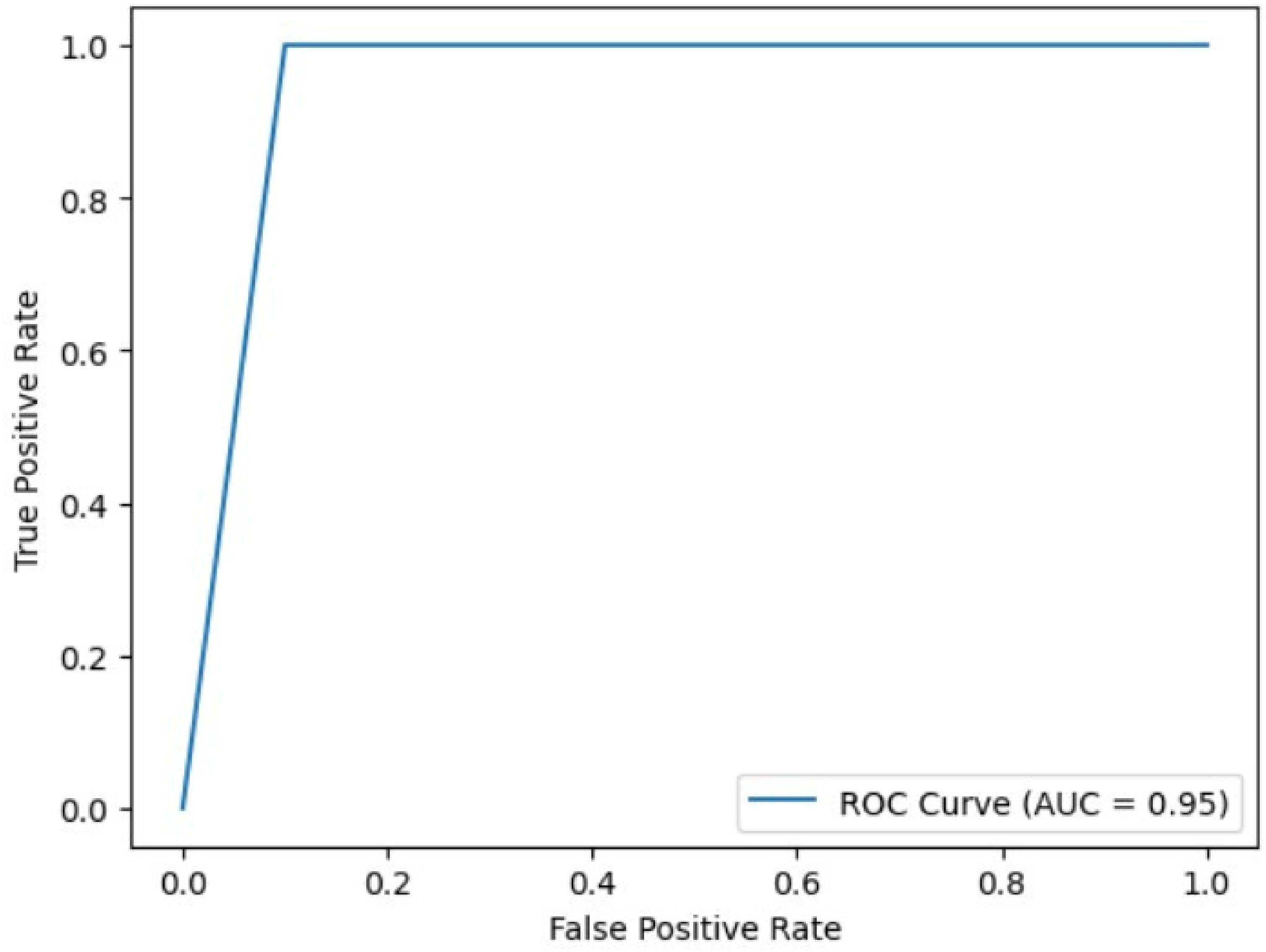
- It is a table that is used to evaluate the performance of a classification model. It shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class in the classification.
- TP and TN are the correct predictions.



GRAPHS AND TABLES

AUC-ROC CURVE :

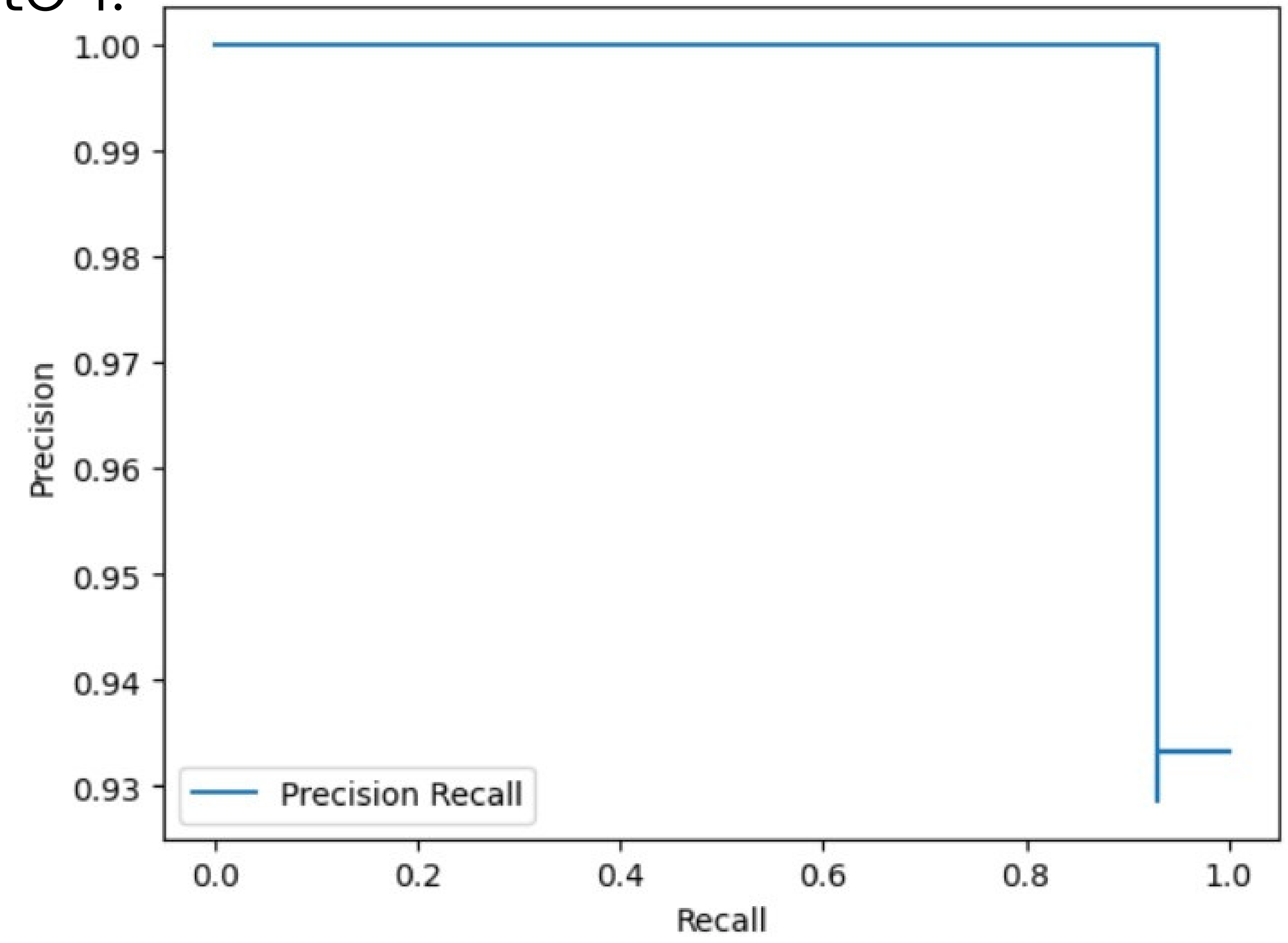
- ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier. It shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various classification thresholds.
- A perfect classifier would have an ROC curve that passes through the top-left corner, with a TPR of 1 and an FPR of 0.
- The area under the curve(AUC) for ROC curve of our model is 0.95.



GRAPHS AND TABLES

AUC-PR CURVE :

- The Precision-Recall (PR) curve is a graphical representation of the performance of a binary classifier. It shows the trade-off between precision and recall at various classification thresholds.
- The PR curve plots precision on the y-axis and recall on the x-axis, and a perfect classifier would have a PR curve that passes through the top-right corner, with both precision and recall equal to 1.
- The area under the curve(AUC) for PR curve of our model is 0.995.



CONCLUSION

- In conclusion, the model that we have built using SVM classifies the given dataset into their respective classes.
- The accuracy and precision scores are also good enough but not the best as no classifier does the exact classification due to outliers.
- We have successfully followed the model architecture and plotted various comparision graphs and tables such as "Confusion Matrix", "AUC-ROC curve" and also "AUC-PR curve".
- Cancer classification is a complex and ongoing process that involves the identification and characterization of different types of cancer based on various factors such as genetic mutations, tissue morphology, and clinical presentation.
- However, cancer classification remains a challenging and evolving field, as new types of cancer are continuously being identified and existing classifications are refined.

REFERENCES

- "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review" - Fadi Alharbi and Aleksandar Vakanski
- "PEM: Accurate cancer type classifcation based on somatic alterations using an ensemble of a random forest and a deep neural network" - Kanggeun Lee, Hyoung-oh Jeong, Semin Lee & Won-Ki Jeong
- "Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations" - Yukun Chen, Jingchun Sun, Liang-Chin Huang, Hua Xu & Zhongming Zhao
- "Evaluating machine learning methodologies for identifcation of cancer driver genes." - Sharaf J. Malebary & Yaser Daanial Khan
- "Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas." - Gregory P. Way, Francisco Sanchez-Vega & Konnor La

THANK YOU

Varshith Alladi - S20200010012

Manjula Naidu - S20200010139

Viswanath Reddy - S20200010086

Seshu Medapi - S20200010124

Hemalatha Remidala - S20200010177