UCD Michael Smurfit Graduate Business School

Digital Transformation in Healthcare

Presented by: Group 15

Ambarish Tirumalai (23201747)
Bishal Ghosh (23200342)
Stebin Sebastian (23200018)
Thapanee Sasuwan (23201498)

Introduction

Irish Healthcare stands at a critical juncture: to serve the people, it needs to integrate the current fragmented system into a unified one. One of the ways that it can be achieved to triage resource allocation through machine learning.

Empowering Stakeholders

- ***Patients:*** Immediate gains in personalised, timely care
- ***Healthcare professionals:*** Enhanced decision-making, decreasing administrative load
- ***Health Service Executive (HSE):*** Optimised resource distribution, improved care quality
- ***Health Insurance Provider:*** *Risk management, Persoanlised insurance plans and faster claims*

Objectives

- **Resource Management:** Implement ML to refine hospital resource allocation, aligning bed availability and staffing to demand
- **Patient Triage:** Integrate AI systems to evaluate initial symptoms, directing patients accurately at the start of their healthcare journey (common diseases)

Study Scope

01

Hospital admission

- Categorisation between OPD and emergency admissions
- Forecasting admission number of patient for next 3 months

02

Chronic diseases

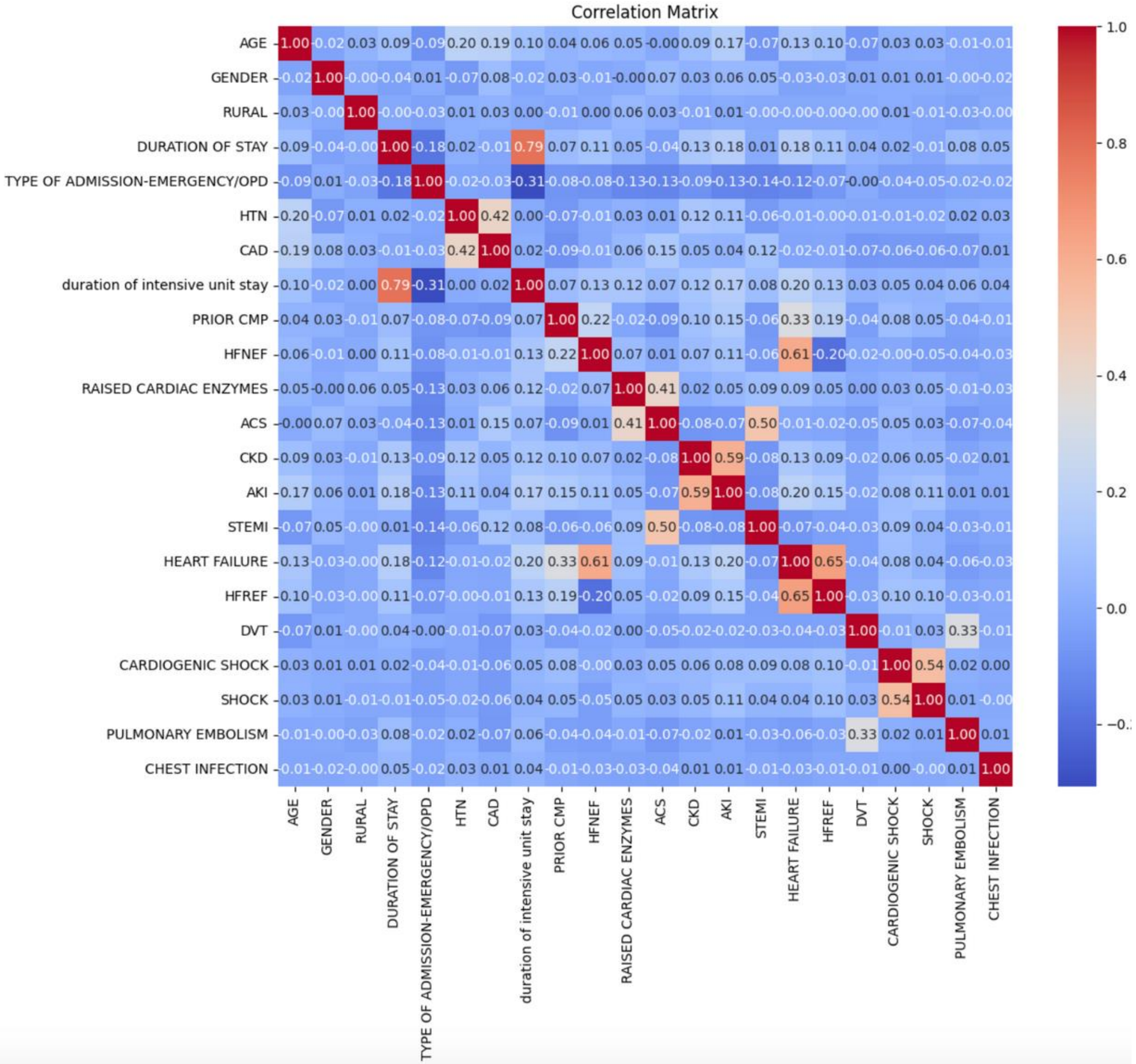
- Diabetes
- Kidney
- Heart

03

Cancers

- Colon
- Leukaemia
- Prostate
- Breast
- Lung

Hospital Admission



Model selection

Logistic Regression: Handles continuous and categorical data, interpretable results (easier to understand which factors influence prediction), computationally efficient.

Random Forest (RF): Combines multiple decision trees, trained on random data subsets with replacement and random feature subsets at splits.

Preprocessing

Using key features, after feature engineering, we used **logistic regression** to determine if a patient's case is Emergency or OPD. Similarly, using multivariate analysis between key features we used Random forest to get estimation of number of admissions for next 3 months.

Summary

The trained model achieved an **accuracy** of **81.22%**, effectively predicting between emergency and OPD cases.

Considering patients with multiple admission, the predicted admission rate for randomly selected **2,000** patients is **84.65%** over the next 3 months. However, the estimated percentage drops to **28.55%** when taking into account the entire unfiltered data.

Chronic diseases - DIABETES

Model selection

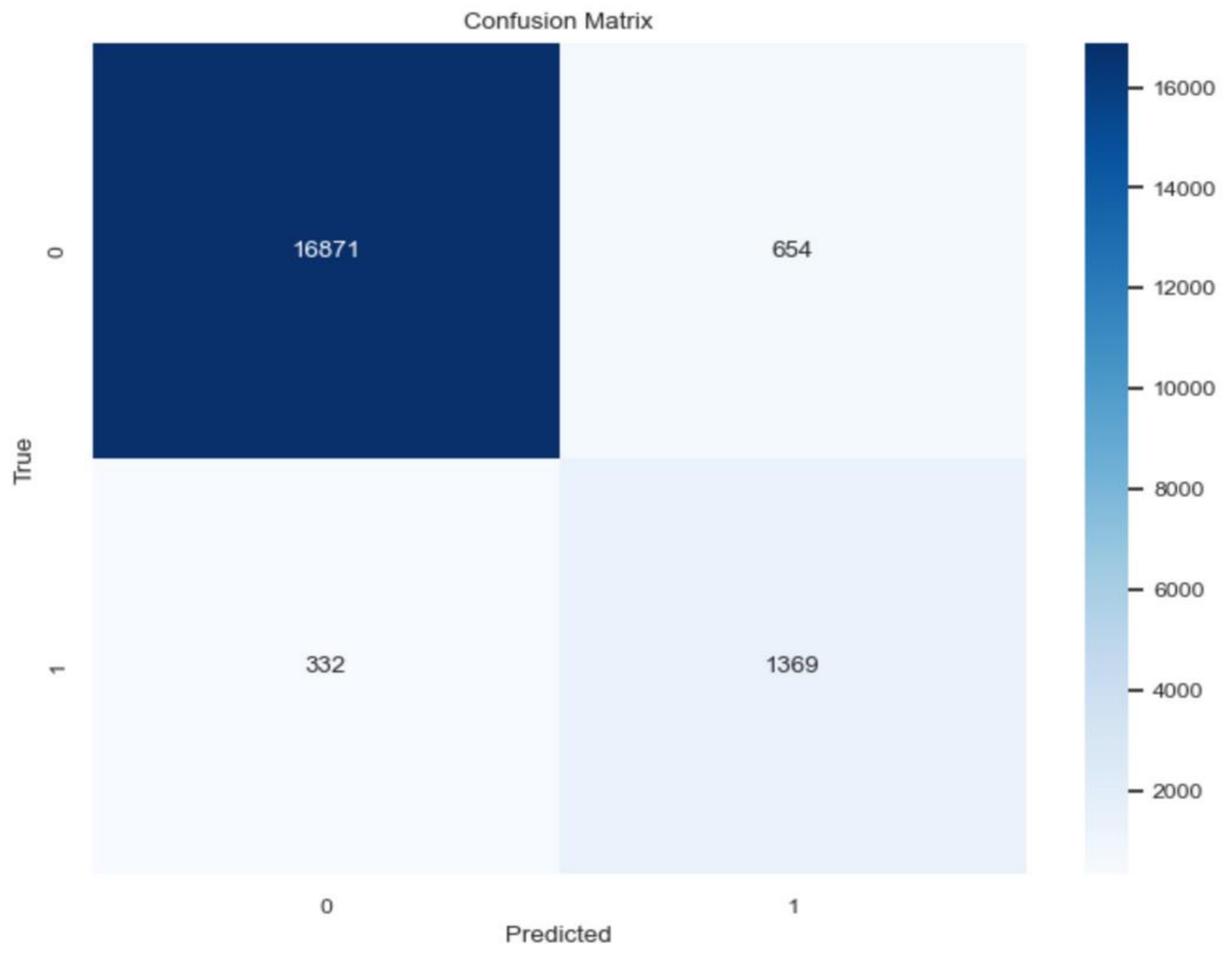
Random Forest combine multiple decision trees, trained on random data subsets with replacement and random feature subsets at splits. Aggregating these trees' predictions improves overall accuracy.

Preprocessing

One-Hot Encoding effectively transforms categorical variables into numerical representations, ensuring the model can understand these features.

Standard Scalar is utilised to normalise all input features, ensuring a mean of 0 and standard deviation of 1. This creates a balanced data scale for machine learning models, regardless of imbalanced datasets.

Grid Search helps to optimise model by efficiently exploring a range of hyperparameter combinations to identify the settings that yield the best performance.



Accuracy	Recall	Precision	F1 score
95%	0.80	0.68	0.74

Summary

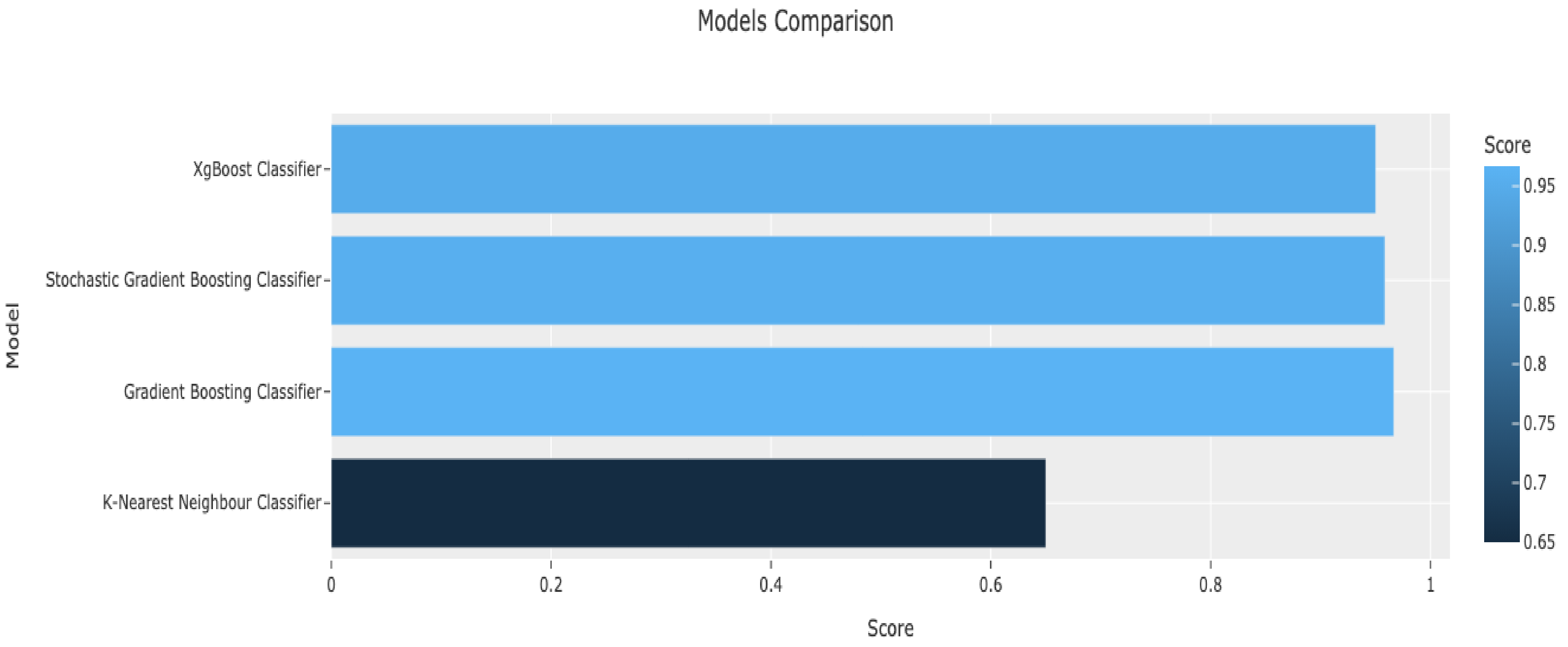
To assess performance, the trained model was evaluated and achieved an accuracy of around **95%.** This indicates the model effectively classified most cases.

The model reveals key factors influencing diabetes in our model. **HbA1c (blood sugar control), blood glucose level, age, and BMI** emerged as the most important features

Chronic diseases –KIDNEY

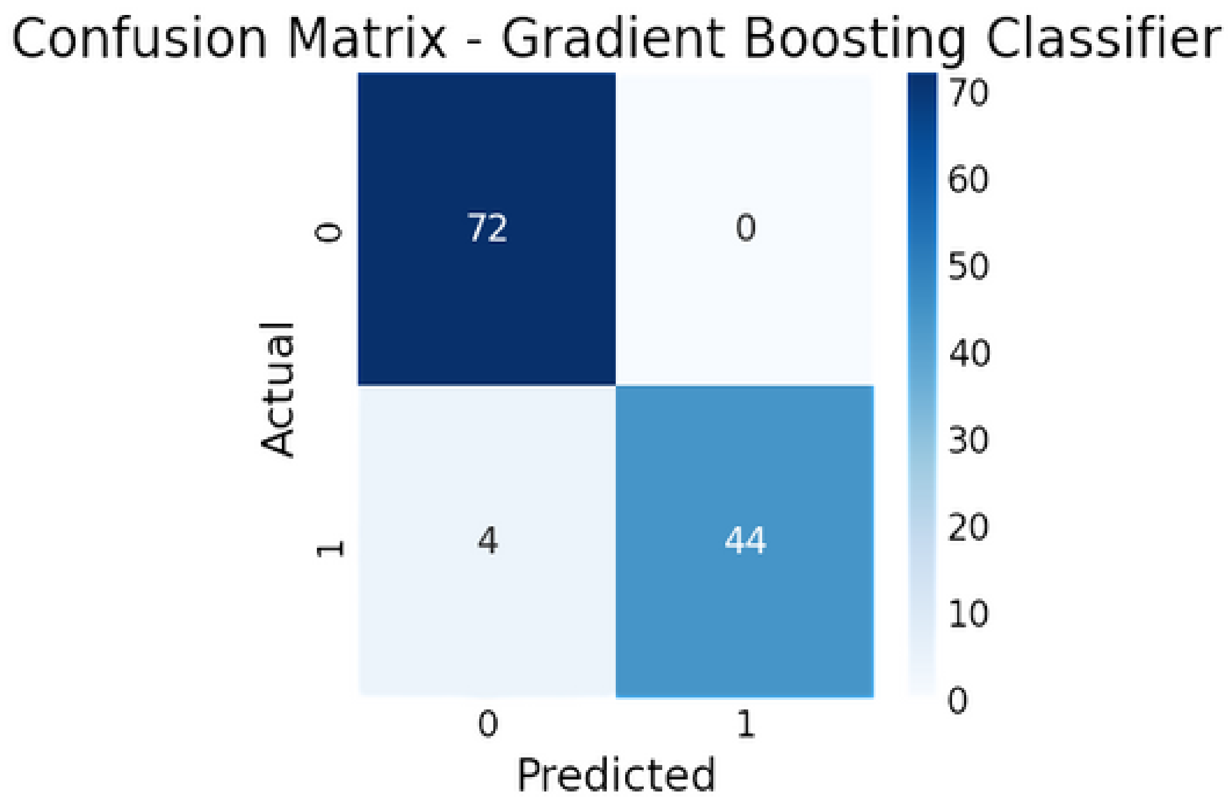
Model selection:

- **Gradient Boosting (GB), XGBoost, Stochastic gradient boosting (SGB):** Powerful for complex, non-linear relationships: Can capture intricate interactions between features that might influence kidney disease development.
- **K-Nearest Neighbors (KNN):** a user-friendly choice for high-dimensional kidney disease prediction due to its minimal parameter tuning and ability to handle datasets with many features.



Preprocessing

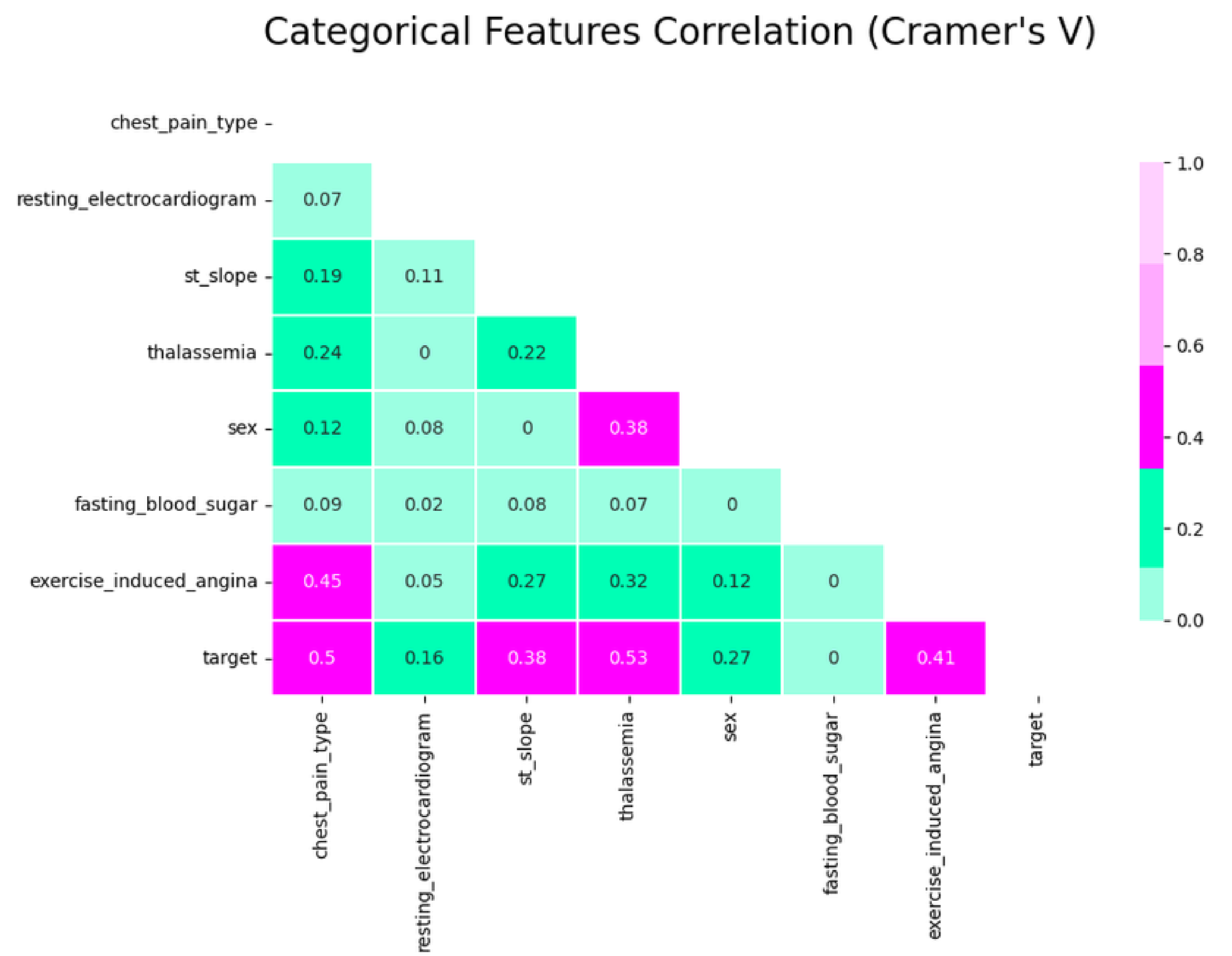
Label Encoder is used transforming categorical data into numerical labels for machine learning algorithms.



Summary

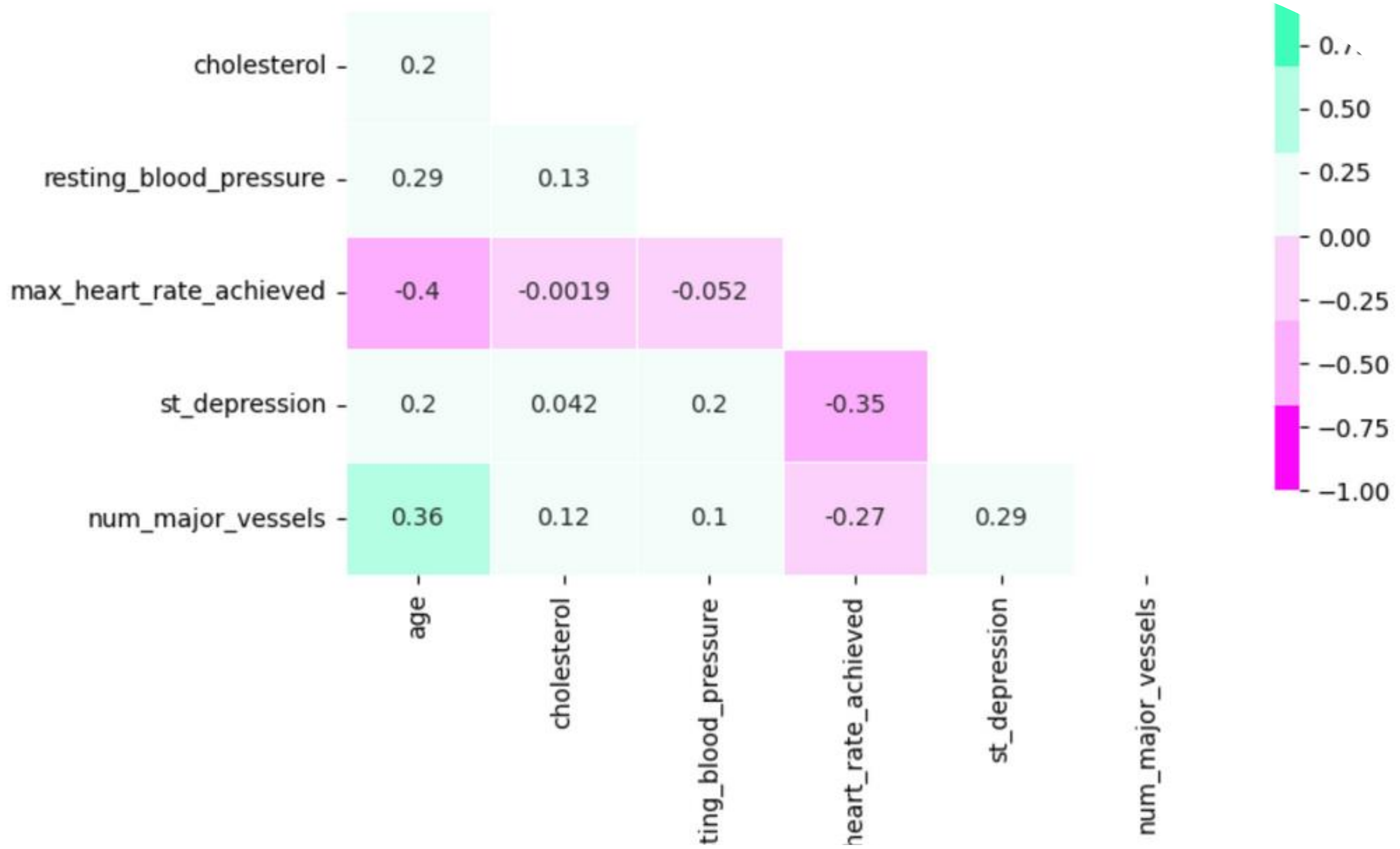
Aside from KNN, most models excelled in kidney disease detection, achieving accuracy above **95%**. Analyzing **precision, recall, and F1-score (all exceeding 90%),** we see strong performance in identifying healthy individuals. This suggests the models correctly classified a high percentage of healthy cases.

Chronic diseases - HEART



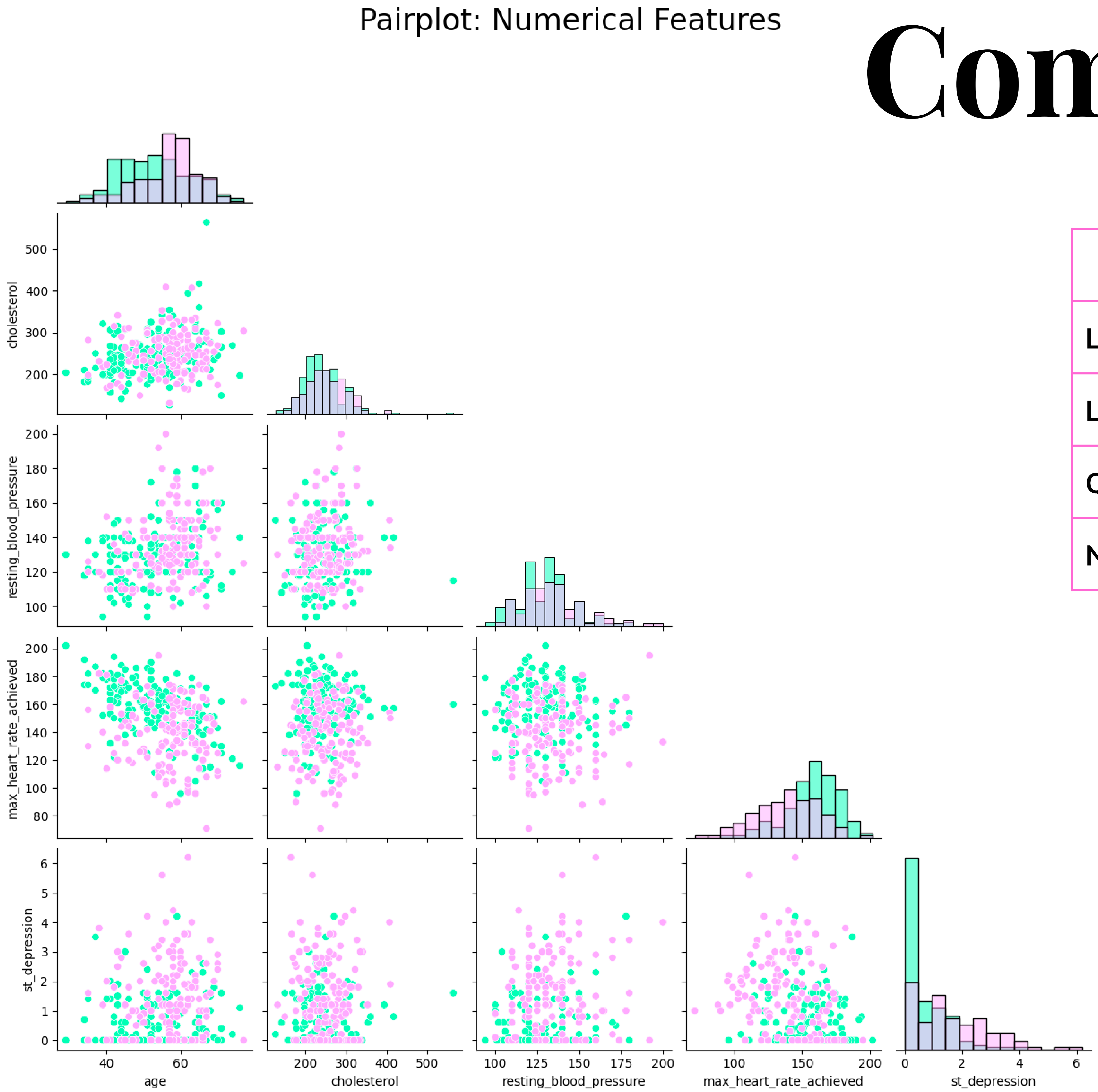
Model selection:

- **Logistic Regression (LR)**
- **Naive Bayes**: Simple and efficient, works well with high-dimensional data (many features).
- **Linear Discrimination Analysis (LDA):** Effective for well-separated classes in data, good interpretability.
- **Quadratic Discrimination Analysis (QDA):** Can capture non-linear relationships between features and heart disease, potentially more accurate than LDA for complex data.



Preprocessing

Label Encoder is used transforming categorical data into numerical labels for machine learning algorithms.



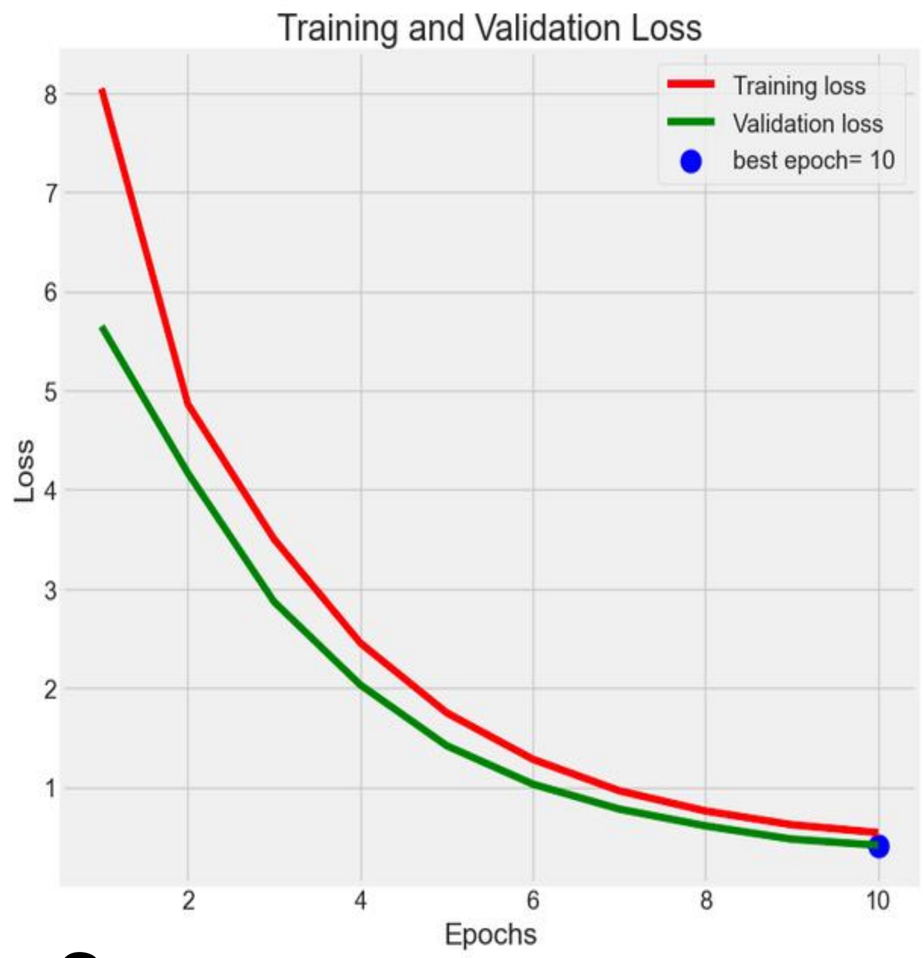
Comparison Result

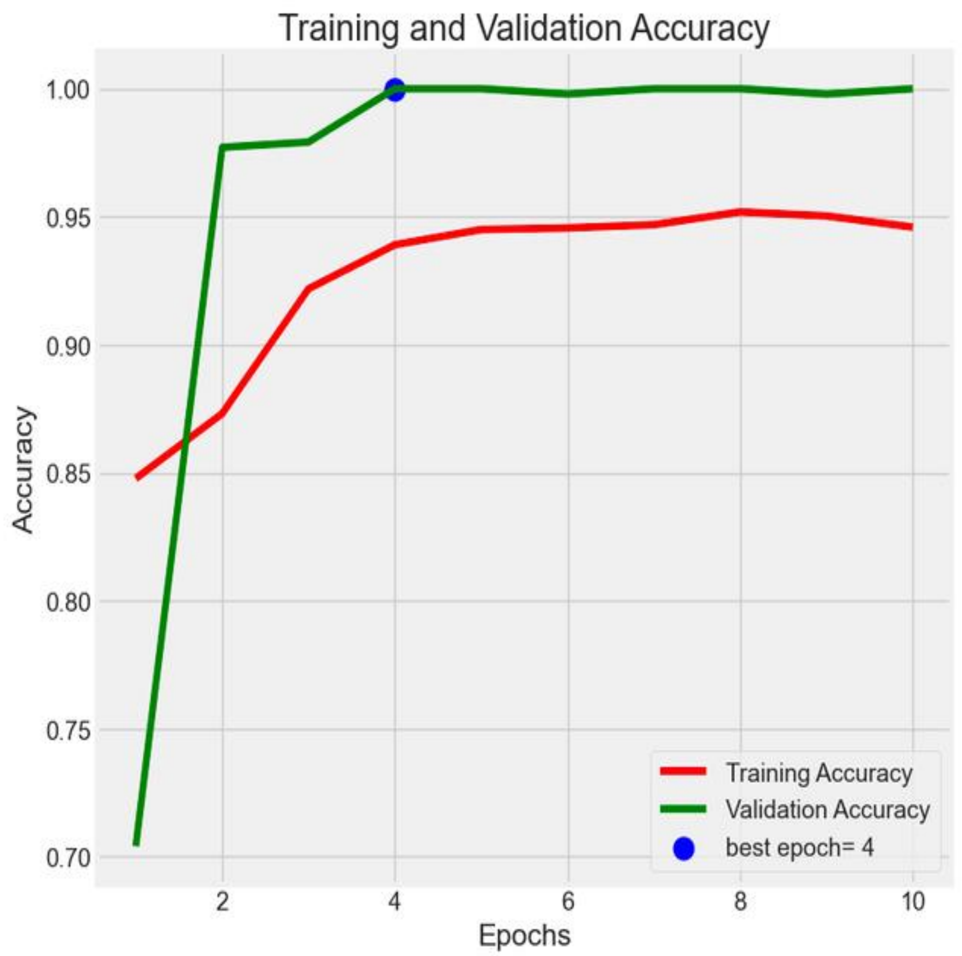
	Accuracy	Recall	Precision	F1 Score
Logistic Regression	**86%**	0.91	0.82	0.86
Linear DA	**85%**	0.89	0.82	0.85
Quadratic DA	**85%**	0.83	0.85	0.84
Naive Bayes	**82%**	0.86	0.79	0.82

Summary

Excluding Naive Bayes, most models achieved high accuracy (**>85%**) in detecting heart disease using features like **chest pain type, major vessels involved, thalassemia, exercise with angina, max heart rate, and ST depression.** Focusing on precision, Logistic Regression and Linear DA excelled with over 89% accuracy, indicating these models effectively ruled out heart disease in a significant portion of healthy cases.

Cancers - COLON CANCER & LEUKAEMIA





Model Selection

EfficientNet - B3 is a convolutional neural network (CNN) built to train models on images, based on the concept called Compound Scaling. This concept addresses the tradeoff between three essential dimensions of a neural network: Width, Depth Resolution

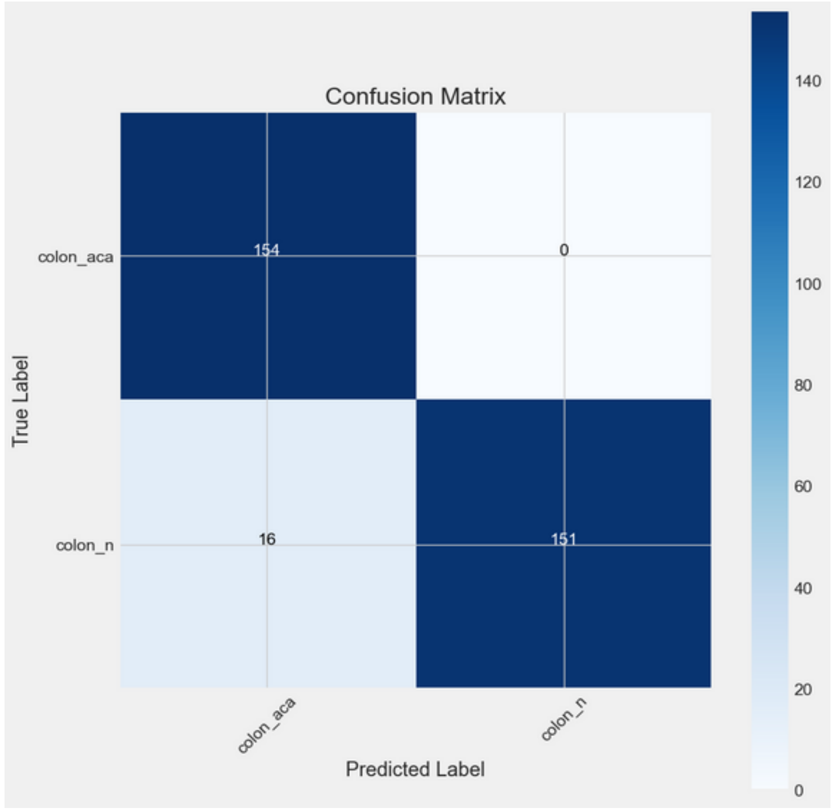
Preprocessing

Adamax is used where the gradients have high variability or sparse features are involved.

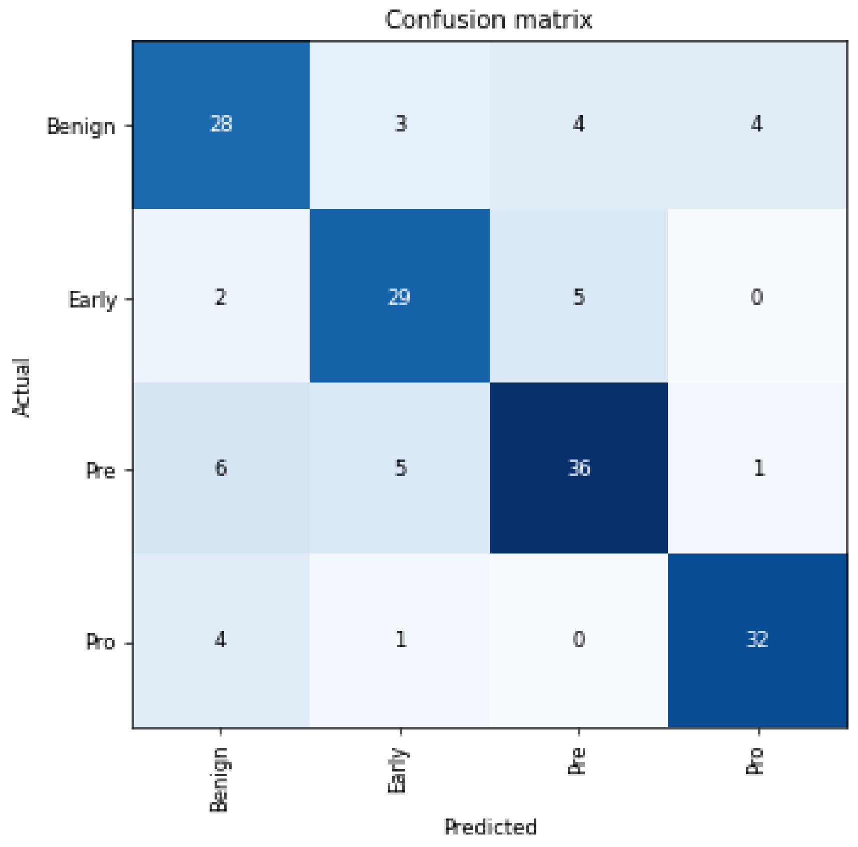
	Accuracy	Recall	Precision	F1 score
Colon Cancer	87.5%	1.0	0.4167	0.5880
Leukaemia	85.71%	1.0	0.33	0.5

Summary

The accuracies of both cancers are more than **85%,** which shows that the EfficientNet - B3 model is very apt for analysing image-based data.



Colon Cancer Confusion Matrix



Leukemia Confusion Matrix

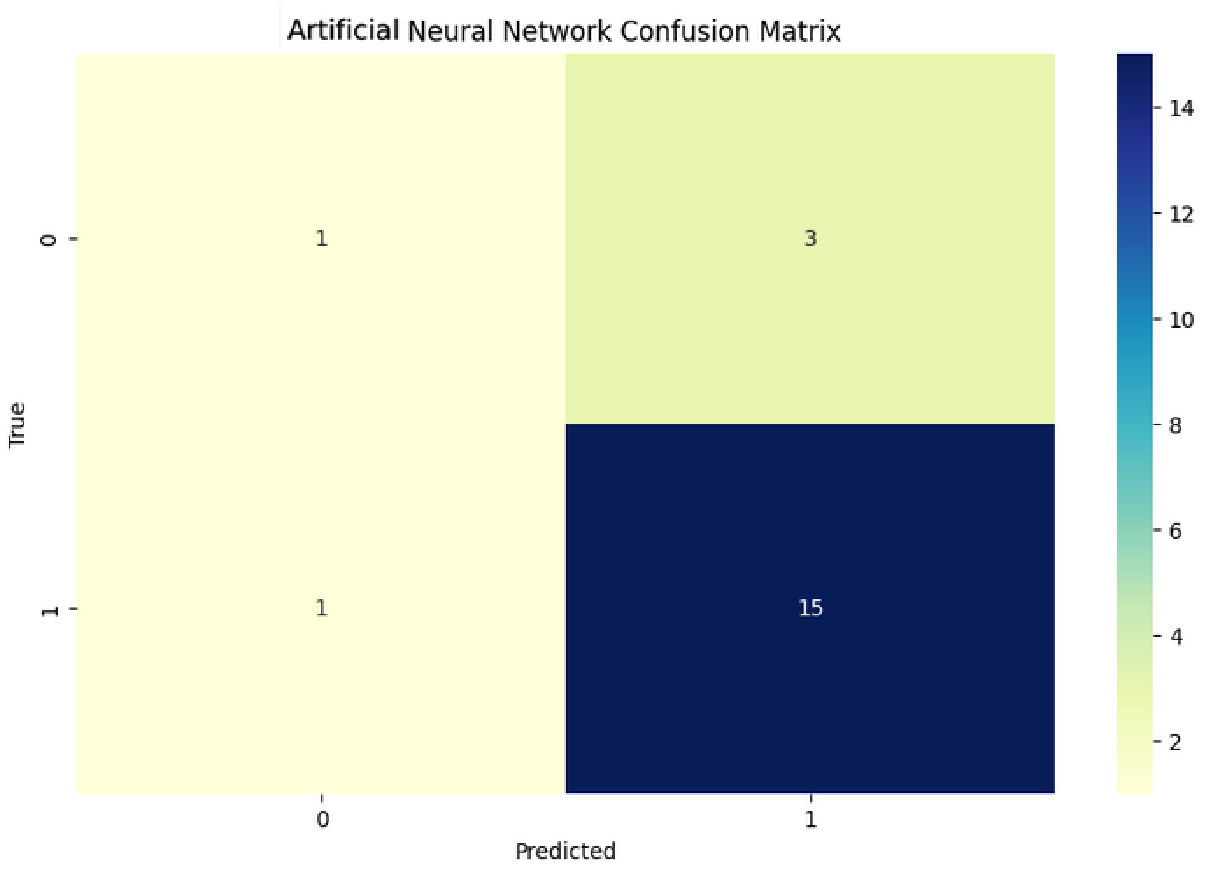
Cancers - PROSTATE CANCER

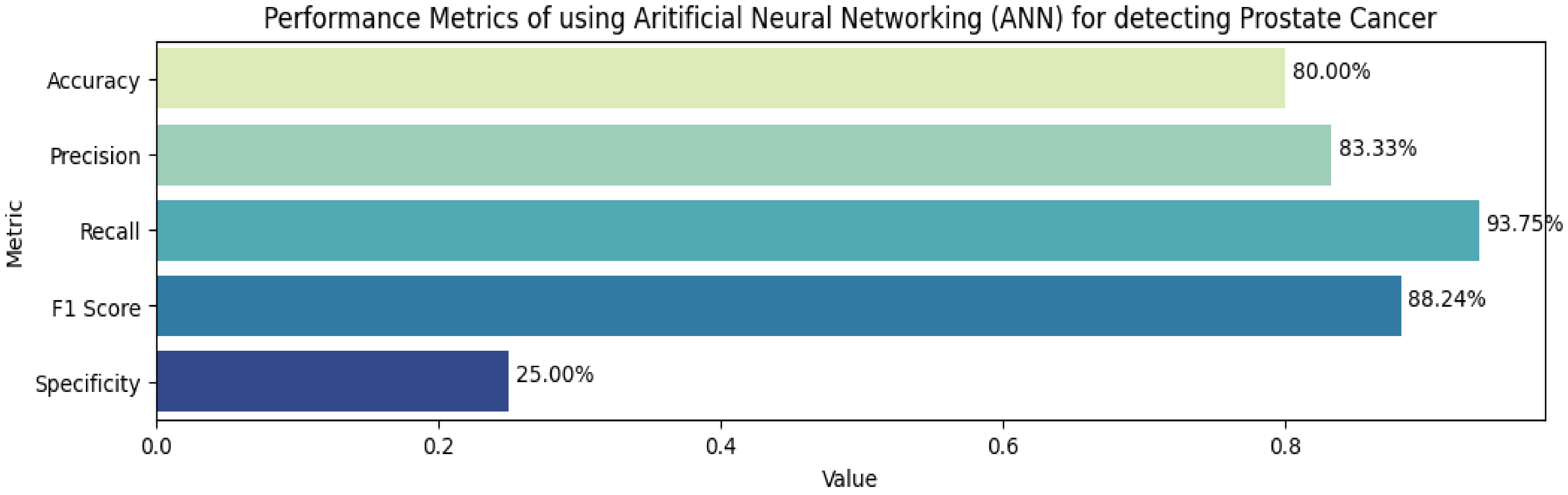
Model Selection

Artificial Neural Networks (ANN) are good at identifying complex patterns and relationships between features, which can then be generalised on an unseen dataset after training the model. It can adapt to the data provided, optimizing the weights to improve performance.

Preprocessing

MinMax Scaler is used where the features are transformed into a given range





Evaluation

To assess performance, the trained model was evaluated on a test set through a **confusion matrix**, achieving an accuracy of around **80%,** and a recall value of around **93.75%.** This indicates the model effectively classified most cases.

Accuracy	Recall	Precision	F1 score
80%	0.9375	0.8333	0.8824

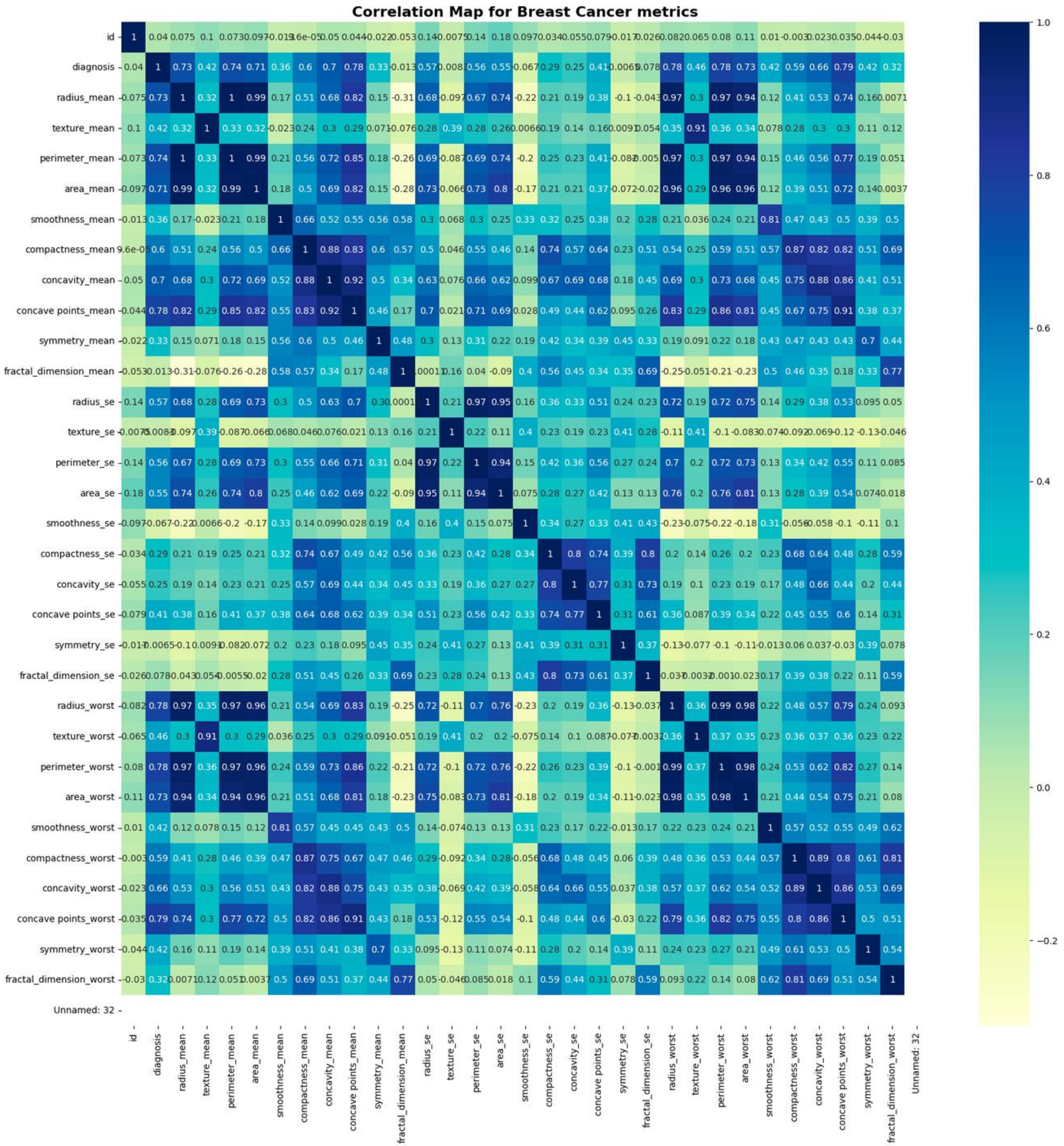
Cancers – BREAST CANCER

Preprocessing

Standard Scaler is used to normalise the data by removing the mean and scaling to unit variance

Model Selection

- **Logistic Regression** is a model designed for binary classification problems, that has a probabilistic interpretation and whose model coefficients can be related to the odds ratios of each predictor. It is computationally simple.
- **Support Vector Machine (SVM)** is a model that finds the largest possible margin that can separate the classes, which results in a good generalisation of test data.

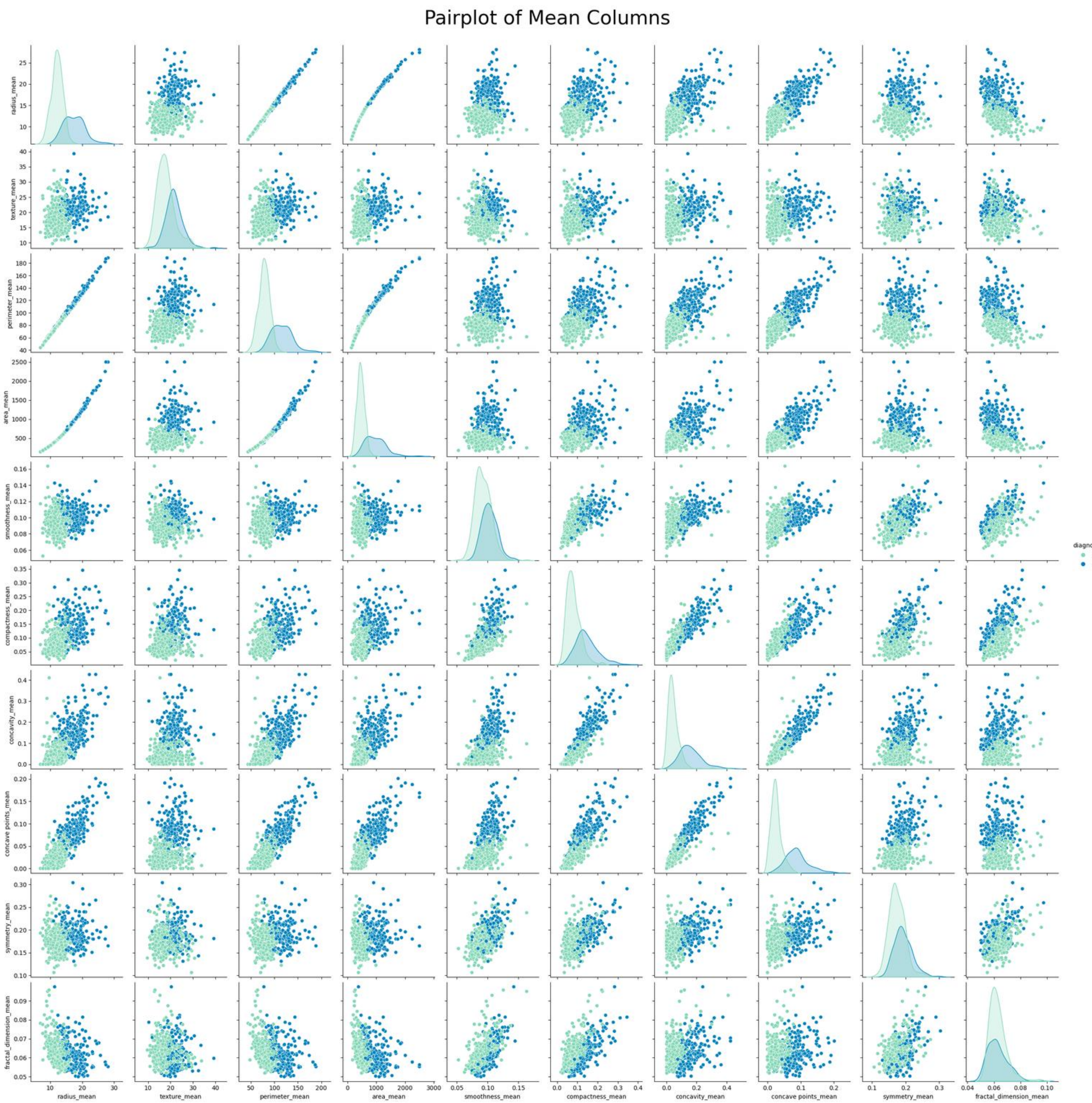


Comparison Result

Summary

We can see from the results that both models work equally well, and both are equally adept at evaluating the possibility of breast cancer

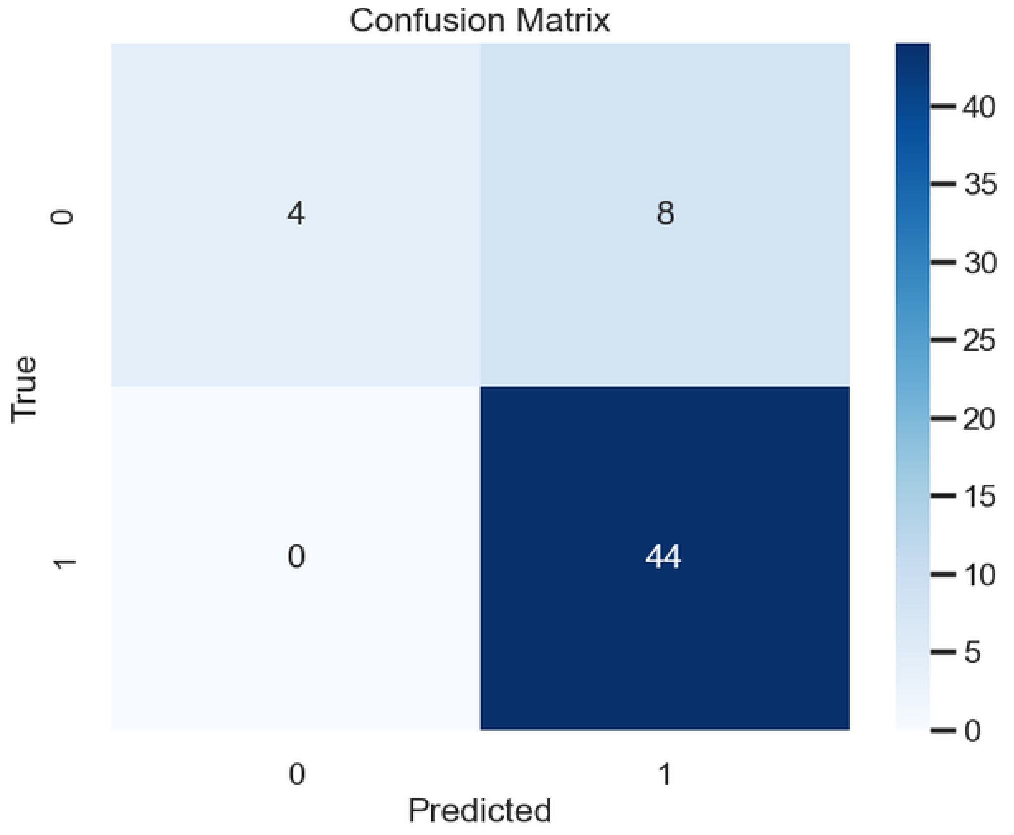
	Accuracy	Recall	Precision	F1 score
Logistic Regression	98.25%	0.9726	1.0	0.9861
SVM	98.25%	0.9726	1.0	0.9861



Cancers – LUNG CANCER

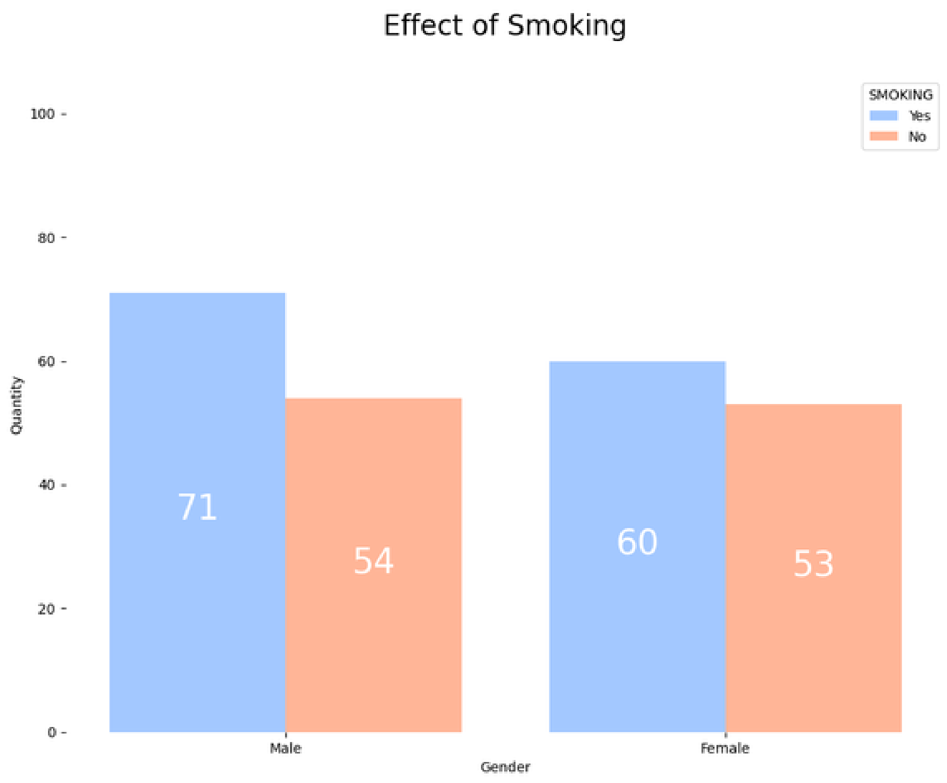
Preprocessing

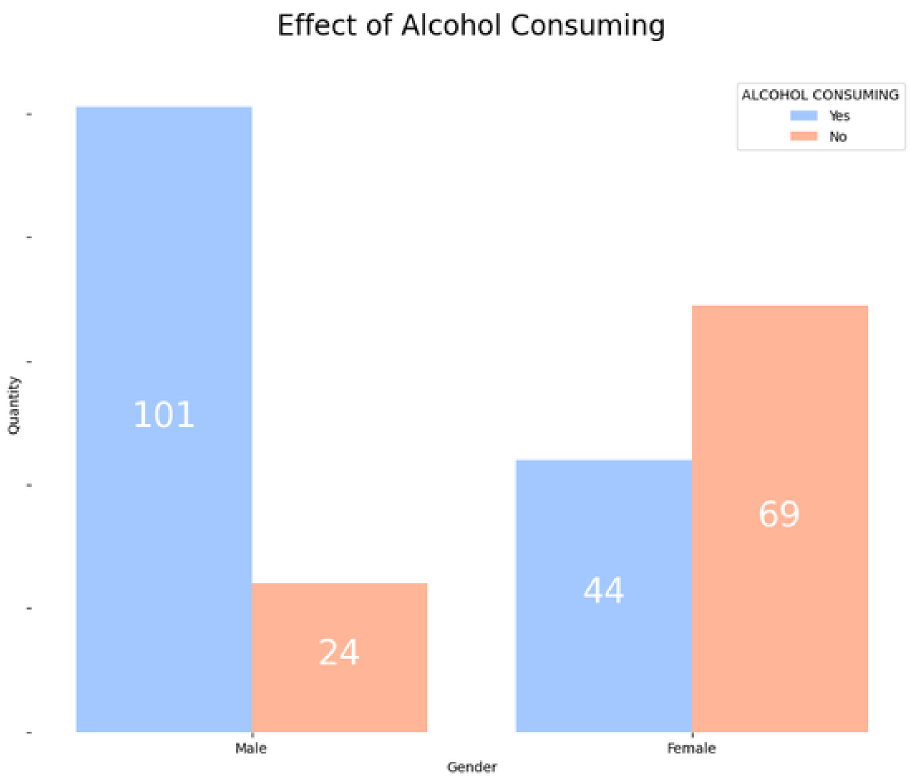
Label Encoder and **Standard Scaler** is used to transform and normalise the data

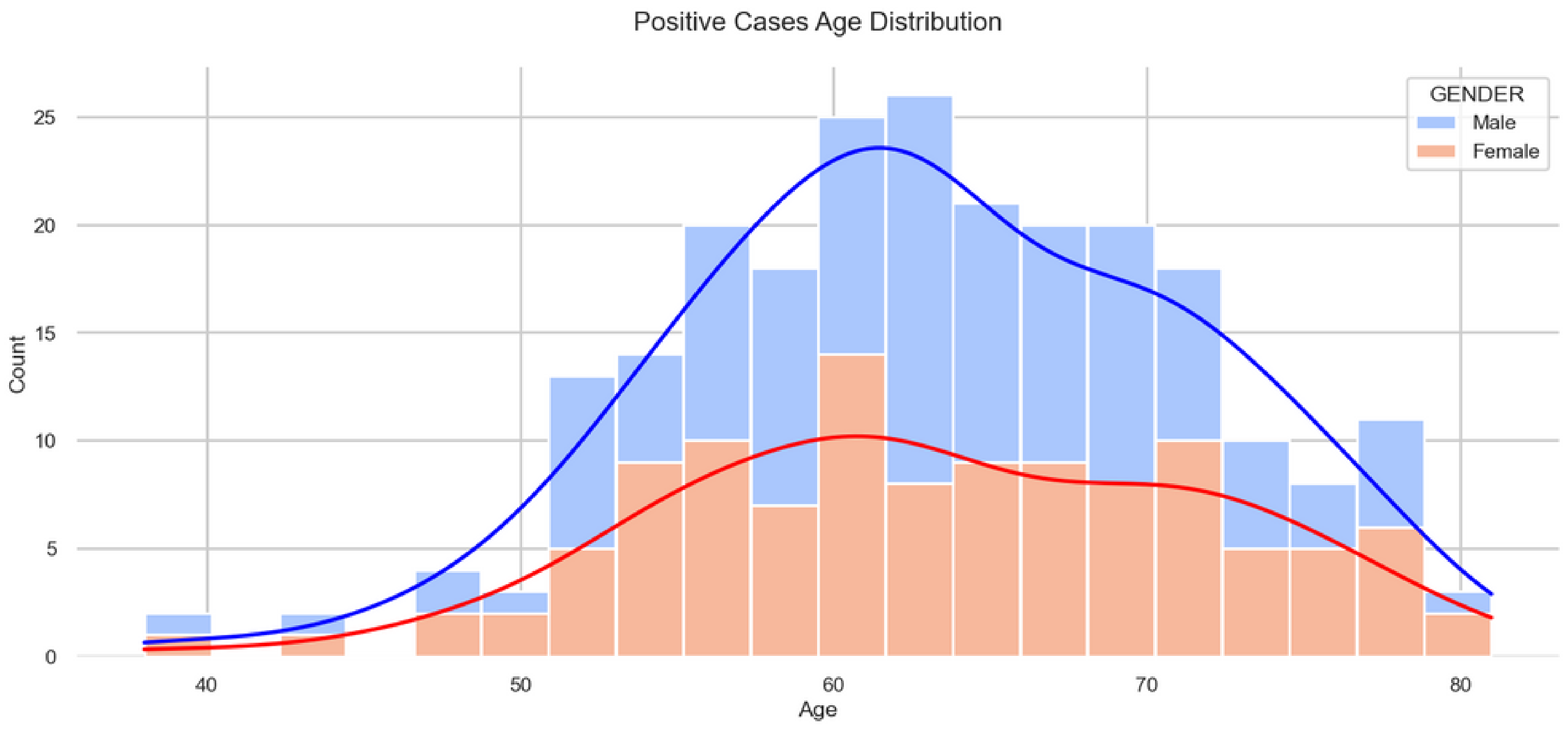


Model Selection

- **Logistic Regression** is a computationally simple model designed for binary classification problems, that has a probabilistic interpretation and whose model coefficients can be related to the odds ratios of each predictor.
- **Support Vector Machine (SVM)** is a model that finds the largest possible margin that can separate the classes, which results in a good generalisation of test data.
- **Neural Network** The neural network used here is called a Forward Feed Neural Network, which has a sequential architecture and no feedback loops, which is why it is highly flexible.







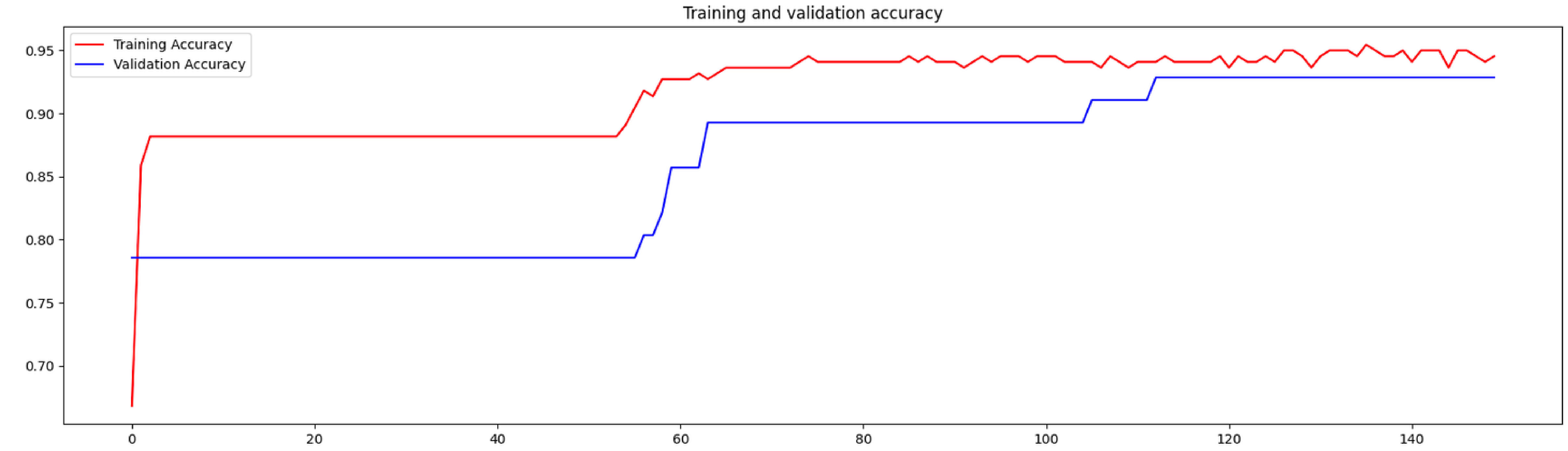


Image: Training and validation accuracy is plotted for 150 epochs

Comparison Result

Summary

From the results, it can be seen that Forward feed neural network is the most accurate in predicting the possibility of lung cancer in patients, with an accuracy value of **92.86%**. It excels in all parameters except the recall factor, which shows the success of the algorithm in this classification.

	Accuracy	Recall	Precision	F1 score
Logistic Regression	88%	1.0	0.4167	0.5880
SVM	86%	1.0	0.33	0.5
Forward Feed NN	approx. 90%	0.9	0.75	0.82

Next steps

- Further research with a larger, Irish-specific patient dataset.
- Collaboration with healthcare domain experts to refine the model.
- Pilot program in a selected hospital to assess real-world performance.
- Address data gaps and potential biases for successful large-scale implementation.

References

[1] Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. Journal of Information System Exploration and Research, 2(1). https://doi.org/10.52465/joiser.v2i1.245

[2] Islam MA, Majumder MZH, Hussein MA. (2023). Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inform. 2023 Jan 12;14:100189. doi: 10.1016/j.jpi.2023.100189.

[3] Barhoom, Ali M. A. et al (2022). Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. International Journal of Engineering and Information Systems (IJEAIS) 6 (4):1-13.

[4] Bemando, C., Miranda, E. & Aryuni, M. (2021), "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms", IEEE, , pp. 232

[5] Young, D.L., Engels, R., Colantuoni, E., Friedman, L.A. & Hoyer, E.H. (2023), "Machine learning prediction of hospital patient need for post-acute care using an admission mobility measure is robust across patient diagnoses", Health policy and technology, vol. 12, no. 2, pp. 100754.

[6] Bollepalli, S.C., Sahani, A.K., Aslam, N., Mohan, B., Kulkarni, K., Goyal, A., Singh, B., Singh, G., Mittal, A., Tandon, R. et al., 2022. An optimized machine learning model accurately predicts in-hospital outcomes at admission to a cardiac unit. Diagnostics, 12(2), p.241. Available at:https://doi.org/10.3390/diagnostics12020241

[7] Ghaderzadeh, M., Aria, M., Hosseini, A., Asadi, F., Bashash, D. and Abolghasemi, H., 2021. A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images. International Journal of Intelligent Systems, [online]

[8] Burrows, N.R., 2017. Incidence of end-stage renal disease attributed to diabetes among persons with diagnosed diabetes—United States and Puerto Rico, 2000–2014. MMWR. Morbidity and mortality weekly report, 66. Available at: https://doi.org/10.24432/C53919

[9] Rubini, L., Soundarapandian, P. and Eswaran, P., 2015. Chronic Kidney Disease. [online] Available at: https://doi.org/10.24432/C5G020

[10] Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R., 1988. Heart Disease. UCI Machine Learning Repository. Available at: Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R., 1988. Heart Disease. [online] Available at: https://doi.org/10.24432/C52P4X

[11] Wolberg, W., Mangasarian, O., Street, N. and Street, W., 1995. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C5DW2B

[12] Saifi, S. and Mahmood, S., 2017. Prostate Cancer. Kaggle. Available at: https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer

[13] Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A. and Mastorides, S.M., 2019. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv preprint arXiv:1912.12142. Available at: https://arxiv.org/abs/1912.12142

[14] Al Aswad, N., 2023. Lung Cancer [Dataset]. Available at: https://huggingface.co/datasets/nateraw/lung-cancer