TRANSACTIONS ON EMERGING SCIENCES
Ideas at the Edge of Discovery
ISSN 3136-2710

ORIGINAL RESEARCH ARTICLE

COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHMS FOR DISEASE PREDICTION

Samreen Fatima
Department of Computer Science,
University of Agriculture, Faisalabad, Pakistan
Corresponding Author Email: mianateeq027@gmail.com

ABSTRACT

Machine learning has become a key tool in healthcare, particularly for the early detection of disease and for supporting the diagnostic process. This research evaluates the comparative performance of six predictive models — Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbours (KNN) — across five medical datasets: Heart Disease, Chronic Kidney Disease (CKD), Breast Cancer Wisconsin Diagnostic (BCWD), Hepatitis C, and Cirrhosis Patient Survival Prediction. Filter-based feature selection techniques — Chi-square, Mutual Information, ANOVA F-test, and Correlation Analysis — were used to enhance prediction performance. Model performance was evaluated using three train-test split ratios (60/40, 70/30, 80/20) and three cross-validation configurations (5-fold, 7-fold, 10-fold), using accuracy, precision, F1-score, and RMSE as evaluation metrics. Results show that Random Forest achieved the best overall performance on most datasets, and Mutual Information was the most effective feature selection method. Averaged across configurations, models achieved an accuracy of approximately 63.40%, precision of 63.40%, and F1-score of about 63.14% on the most challenging dataset, with an RMSE around 0.96, while performance on the remaining datasets was substantially higher and stable across validation splits. Classification results were near-perfect for Chronic Kidney Disease but consistently most difficult for Heart Disease. These findings underline the importance of feature selection and rigorous comparative evaluation for building effective and reliable disease-prediction models in healthcare.

Keywords: *Machine Learning; Disease Prediction; Feature Selection; Random Forest; Cross-Validation; Medical Datasets; Comparative Analysis*

1. INTRODUCTION

Over the past few years, the deployment of machine learning in the healthcare domain has attracted considerable attention due to its capacity to support early detection and clinical decision-making (Obermeyer and Emanuel, 2016). The growing availability of medical datasets and computational resources has enabled researchers to develop predictive models capable of analyzing complex patterns in patient data. Early and accurate prediction of diseases such as chronic kidney disease, heart disease, breast cancer, hepatitis C, and liver cirrhosis is important, as it can contribute to better outcomes, lower mortality, and more efficient use of healthcare resources (Kononenko, 2001). Diagnostic approaches are typically based largely on clinical experience, laboratory tests, and imaging techniques, which can be time-consuming, expensive, and prone to manual error. Machine learning offers an alternative approach that builds on historical patient information to

ISSN 3136-2710

produce predictive models that can identify disease patterns with high accuracy (Das et al., 2009). These models can support healthcare workers by offering decision-support systems that improve diagnostic efficiency and reliability in clinical practice.

The performance of machine learning models is highly sensitive to several factors, including data integrity, feature reliability, model selection, and evaluation strategy (Chandrashekar and Sahin, 2014). Because of variation across datasets, algorithms can behave differently depending on data distribution, feature characteristics, and noise levels. It is therefore important to perform a comparative evaluation of multiple machine learning algorithms to determine their effectiveness across different disease-prediction scenarios. While previous studies have generally examined a single classifier or a specific dataset in isolation, a systematic comparative evaluation spanning six classifiers, five medical datasets, four feature selection methods, three split ratios, and three k-fold configurations within a standardized framework remains underexplored. The present research is intended to fill this gap.

Feature selection is another important part of predictive modelling. Medical datasets often contain redundant or irrelevant features that can negatively affect model predictivity, increase model complexity, and lead to overfitting (Liu and Motoda, 2010). Feature selection techniques help identify the most important features contributing to accurate predictions, making models more efficient, accurate, and interpretable. This research conducts an in-depth comparative study of several machine learning algorithms tested on five medical datasets — Chronic Kidney Disease, Heart Disease, Breast Cancer Wisconsin Diagnostic, Hepatitis C, and Cirrhosis Patient Survival Prediction — to provide a more accurate prediction of patient outcomes.

The research focuses on how well each model performs, evaluating it using several metrics including accuracy, precision, F1-score, and root mean square error (RMSE). In addition, the analysis examines the effect of different feature selection methods and how they contribute to predictive performance.

The proposed approach is built around a structured pipeline comprising data analysis, feature selection, model training, model evaluation, and visualisation. To ensure robustness and generalisability of the results, this study uses multiple train-test splits and cross-validation techniques. The findings from this study are intended to provide insight into the best machine learning methods for disease prediction and to highlight the impact of feature selection in healthcare analytics.

1.1 Overview of the Proposed Research Methodology

This study applies machine learning techniques for disease prediction using multiple medical datasets (Kononenko, 2001). Prediction accuracy was evaluated using several classification algorithms, including Decision Tree, SVM, Random Forest, Naive Bayes, Logistic Regression, and KNN. Before model training, data preprocessing and normalisation techniques were applied to improve data quality and model efficiency (Wang et al., 2020). Various feature selection methods — Mutual Information, ANOVA, and Correlation-Based Selection — were used to identify the most relevant features and remove redundant ones (Bashir et al., 2019). The selected features were then used to train and test the models under three train-test split ratios: 60/40, 70/30, and 80/20. Model performance was evaluated using accuracy, precision, F1-score, and RMSE. In addition, 5-fold, 7-fold, and 10-fold cross-validation were used to obtain stable and reliable results (Liu

ISSN 3136-2710

and Motoda, 2010). The overall methodology aims to compare the effectiveness of different machine learning algorithms and feature selection methods for accurate disease prediction.

2. METHODOLOGY

2.1 Overview of Methodology

The research methodology was designed for a systematic assessment and comparison of several machine learning algorithms across different medical datasets. The entire process follows a well-organized pipeline comprising data preprocessing, feature selection, model development, evaluation, and visualisation.

Raw data is first preprocessed to ensure data quality and consistency, including handling missing values and encoding categorical data. Numerical features are scaled using the scikit-learn library (Pedregosa et al., 2011). Feature selection techniques are then applied to identify the most appropriate features, which are used to train several machine learning models. Multiple evaluation strategies are used to assess the models, and results are presented graphically and numerically to aid comparison and interpretation.

2.2 System Architecture and Workflow

The system architecture represents the overall pipeline used in this study, from data input to visualisation of results. The workflow consists of a series of steps, each contributing to improved model performance and reliability.

The process begins with raw data obtained from publicly available medical sources (Dua and Graff, 2019). This data passes through a preprocessing stage that cleans and transforms it into a suitable format. Feature selection techniques are then applied to reduce dimensionality and improve efficiency. The processed data is split into training and testing sets using different splitting ratios, and cross-validation techniques are applied for robustness. Finally, the evaluation results are organised, visualised, and compared using graphical tools.

ISSN 3136-2710

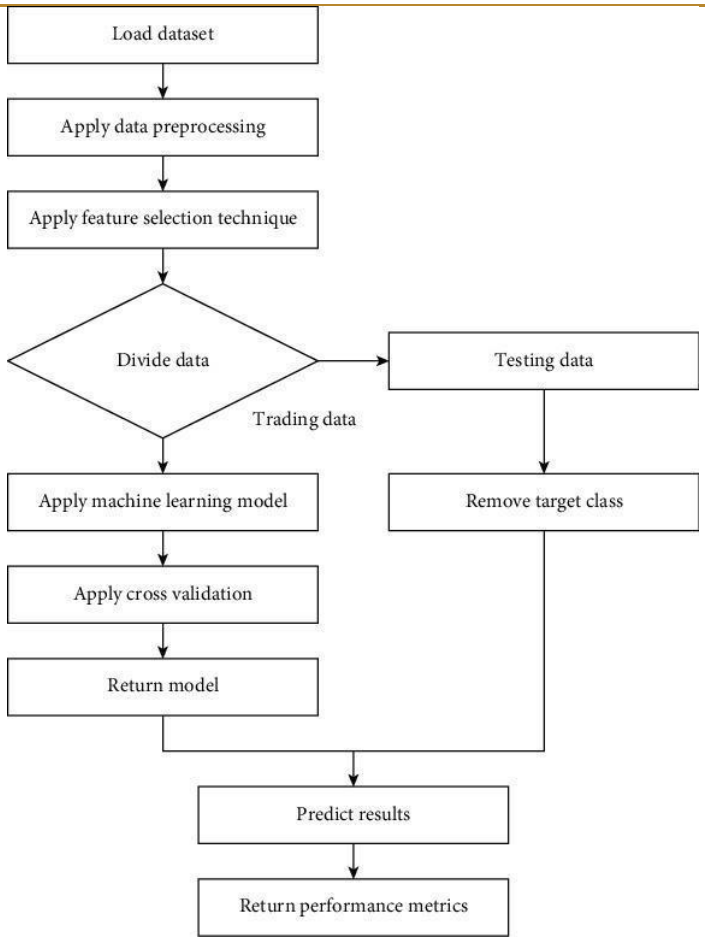

Figure 1. Machine learning disease prediction pipeline (Source: Nararamparambath et al., 2026).

2.3 Datasets

Five publicly accessible medical datasets were used in this study, sourced from the UCI Machine Learning Repository (Dua and Graff, 2019):

1. Breast Cancer Wisconsin (Diagnostic) — n = 569 instances, 30 features, binary classification (malignant / benign).
2. Chronic Kidney Disease (CKD) — n = 400 instances, 24 features, binary classification (CKD / not CKD).
3. Hepatitis C — n = 615 instances, 12 features, binary classification.
4. Cirrhosis Patient Survival Prediction — n = 418 instances, 17 features, survival classification.
5. Heart Disease (Cleveland) — n = 303 instances, 13 features, binary classification (disease present / absent).

2.4 Data Preprocessing

Data preprocessing is an important step in machine learning, as it directly affects model performance (Pedregosa et al., 2011). In this study, several preprocessing techniques were applied to ensure the quality and consistency of the datasets.

ISSN 3136-2710

Missing values were handled using appropriate imputation approaches: numerical features with missing values were replaced using mean imputation, while categorical features were handled using mode imputation. This ensures that no data samples are discarded unnecessarily, preserving dataset size.

Categorical variables were encoded into numerical format using label encoding where categories have an inherent order, and one-hot encoding for nominal variables, avoiding the introduction of unintended relationships.

Feature scaling was performed to standardise the range of numerical variables. Standardisation transforms features to have zero mean and unit variance, while normalisation scales values within a fixed range, typically between 0 and 1. This process is particularly critical for algorithms such as KNN and SVM, where feature scale can substantially affect performance (Akay, 2009).

2.5 Train-Test Split Strategy

The datasets were split into training and testing sets using multiple ratios, including 60/40, 70/30, and 80/20 (Arlot and Celisse, 2010). A larger training set (e.g., 80/20) provides more information for the model to learn from, potentially improving accuracy, while a smaller training set (e.g., 60/40) tests the model's ability to generalise from limited data. This approach helps identify models that remain stable across varying amounts of training data and reduces bias in the evaluation.

2.6 Cross-Validation

Cross-validation was used to ensure the reliability and generalisability of the learning algorithms, following the approach proposed by Kohavi (1995). This study applies k-fold cross-validation with k values of 5, 7, and 10. In k-fold cross-validation, the dataset is split into k parts of equal size; the model is trained on $k - 1$ parts and tested on the remaining part. This process is repeated k times, with each part used as the test set exactly once, and results are averaged across all iterations to obtain the final performance estimate. Using multiple values of k provides a more stable evaluation by reducing variance and improving consistency of the model assessment across different data subsets.

2.7 Feature Selection Methods

Feature selection is applied to enhance model performance by removing irrelevant and redundant features (Chandrashekar and Sahin, 2014). In this study, four filter-based feature selection methods are used:

- Correlation Analysis: identifies features with strong linear relationships with the target variable.
- ANOVA F-Test: analyzes the statistical significance of features by comparing variance between groups.
- Chi-Square Test: measures how strongly categorical variables are related to the target variable.
- Mutual Information: captures non-linear statistical dependencies between features and the target (Liu and Motoda, 2010).

ISSN 3136-2710

Feature selection is performed using two approaches: selection by number (top 5, 10, or 15 features) and selection by percentage (top 10% or 20%). This process reduces dimensionality, improves computational efficiency, and improves model accuracy.

2.8 Machine Learning Models

Several machine learning algorithms were implemented in this study (Pedregosa et al., 2011):

- Support Vector Machine (SVM): finds the optimal hyperplane to separate data points into different classes, particularly effective in high-dimensional spaces (Akay, 2009).
- Random Forest: an ensemble method combining multiple decision trees to increase accuracy and reduce overfitting (Breiman, 2001).
- Decision Tree: a simple, interpretable model that splits data according to feature values.
- K-Nearest Neighbours (KNN): a distance-based algorithm that classifies data points by majority vote among nearest neighbours.
- Naive Bayes: a probability-based classifier that applies Bayes' theorem under the assumption that features are independent (Rish, 2001).
- Logistic Regression: a classification model that predicts class probabilities using the logistic function.

2.9 Evaluation Metrics

Model performance was evaluated using multiple metrics (He and Garcia, 2009):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$F_1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

$$RMSE = \sqrt{[(1/n) \Sigma (y_i - \hat{y}_i)^2]} \quad (4)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives; and y_i and $\hat{y}_i$ are the true and predicted labels for observation i.

2.10 Experimental Setup

The experiments were conducted using Python 3.9 in Google Colab. Libraries used include Pandas for data manipulation, NumPy for numerical computation, and Scikit-learn v1.2.0 for machine learning algorithms (Pedregosa et al., 2011). Results are reported as averages across five independent runs. Visualisation of results was performed using Tableau.

2.11 Visualisation Approach

ISSN 3136-2710

Visualisation plays an important role in interpreting results. Graphical representations such as bar charts, line graphs, and comparison plots are used to analyse algorithm performance across datasets. These visualisations help identify trends, compare evaluation metrics, and understand the impact of feature selection methods.

3. RESULTS

This section presents the experimental results obtained from evaluating six machine learning classifiers — Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF), and Naive Bayes (NB) — across the five medical datasets described in Section 2. Model performance was investigated under three train-test split ratios (60/40, 70/30, 80/20) and three k-fold cross-validation configurations (5-fold, 7-fold, 10-fold). Four filter-based feature selection methods were deployed: Chi-Square (Chi2), Mutual Information (MI), ANOVA F-Test (ANOVA), and Correlation Analysis (Corr). All results are reported as averages across five independent runs.

3.1 Overall Model Accuracy Under the 60/40 Train-Test Split

Table 1 presents the average accuracy (%) of all six classifiers under the 60/40 split using the best-performing feature selection configuration per classifier-dataset pair. Figure 2 provides the corresponding visual overview.

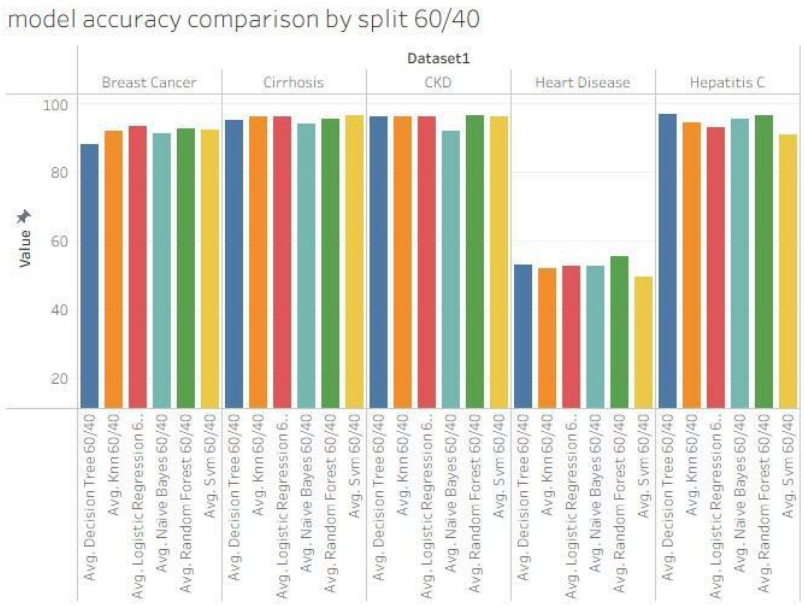

Figure 2. Model accuracy comparison across five medical datasets under the 60/40 train-test split. Each group of bars represents one dataset; individual bars correspond to the six classifiers evaluated. The Heart Disease dataset shows a pronounced performance dip relative to all other datasets.

Dataset	DT	SVM	LR	KNN	RF	NB
Breast Cancer	92.98	97.81	97.37	96.93	96.93	94.74

ISSN 3136-2710

Cirrhosis	95.83	97.02	96.43	97.02	97.02	95.24
CKD	99.38	99.38	99.38	98.75	100.00	93.13
Heart Disease	59.24	62.23	61.96	61.14	62.77	59.78
Hepatitis C	100.00	94.92	100.00	99.58	100.00	99.58

Table 1. Average classification accuracy (%) across datasets under the 60/40 train-test split using the optimal feature selection configuration. DT = Decision Tree; SVM = Support Vector Machine; LR = Logistic Regression; KNN = K-Nearest Neighbours; RF = Random Forest; NB = Naive Bayes.

A clear performance stratification is evident across datasets. The CKD and Hepatitis C datasets consistently obtained the highest accuracy values, with Random Forest and Decision Tree reaching perfect classification (100.00%) under optimal feature configurations. These results suggest that, following rigorous preprocessing and targeted feature selection, the biochemical markers in these datasets contribute exceptionally strong discriminative signals (Sinha and Sinha, 2015; El-Hasnony et al., 2020). Conversely, the Heart Disease dataset exhibited the lowest accuracy across all classifiers (range: 54.35%–62.77%), illustrating the substantially overlapping feature distributions between disease-positive and disease-negative populations (Das et al., 2009). The performance of Naive Bayes was consistently the lowest, particularly on CKD (93.13%), consistent with its known sensitivity to violations of the feature independence assumption (Rish, 2001).

3.2 Dataset Performance Comparison (60/40 Split)

Figure 3 presents a grouped bar chart illustrating the side-by-side accuracy of all six classifiers across each of the five datasets under the 60/40 split.

ISSN 3136-2710

Figure 3. Dataset performance comparison under the 60/40 train-test split. Grouped bars show the accuracy of each classifier per dataset, allowing direct intra-dataset and cross-dataset comparison. Note the tight clustering of bars for Cirrhosis and the substantially lower values for all classifiers on Heart Disease.

For Breast Cancer, values ranged from the lowest (DT: 92.98%) to the highest (SVM: 97.81%), a difference of about 5 percentage points, consistent with previous benchmarks (Mangasarian et al., 1995). The Cirrhosis dataset showed a relatively low standard deviation, with all classifiers clustered tightly together between 95.24% and 97.02%, consistent with Wang et al. (2020). For CKD, five of the six classifiers exceeded 98% accuracy, confirming the strong discriminatory properties of renal biomarkers such as serum creatinine and albumin values (Sinha and Sinha, 2015). The uniformly low accuracy on the Heart Disease dataset (54%–63%) appears to be a property of the dataset itself, with ensemble methods (RF: 62.77%) and margin-based classifiers (SVM: 62.23%) performing slightly better than probabilistic models (NB: 59.78%) (Bashir et al., 2019).

3.3 Split vs. K-Fold Cross-Validation Comparison (Decision Tree)

Figure 4 presents a line graph comparing Decision Tree accuracy across all six evaluation configurations. Table 2 gives the corresponding numerical values.

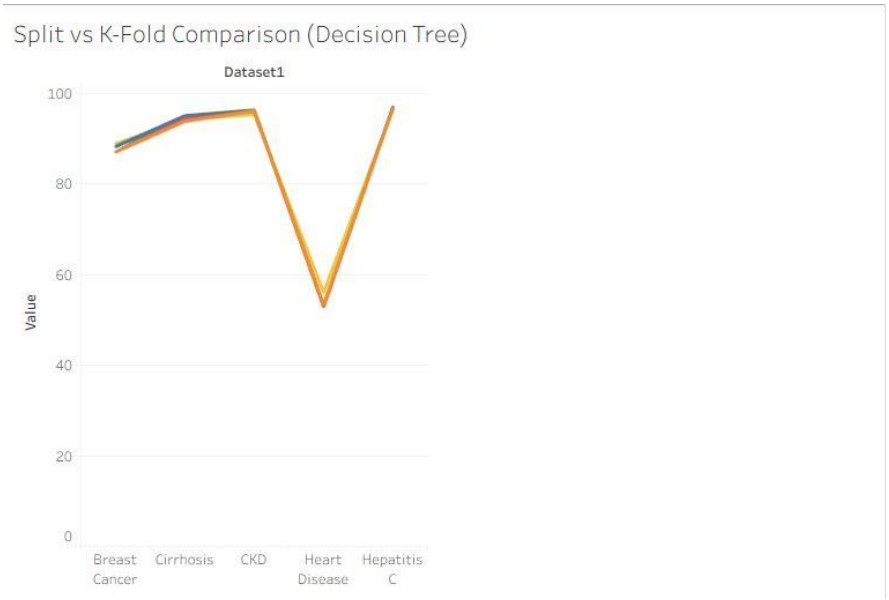

Figure 4. Split vs. k-fold cross-validation comparison for the Decision Tree classifier across five medical datasets. The two lines show the train-test split (solid) and k-fold cross-validation (dashed) strategies. A characteristic dip is observed for Heart Disease under both evaluation approaches, confirming that the low performance on this dataset is an intrinsic property of the data rather than an artefact of evaluation design (Kohavi, 1995).

Dataset	60/40	70/30	80/20	5-Fold	7-Fold	10-Fold
Breast Cancer	92.98	92.40	93.86	91.74	91.74	92.45
Cirrhosis	95.83	95.24	95.24	95.60	95.22	95.60

ISSN 3136-2710

CKD	99.38	99.17	100.00	98.75	98.50	98.75
Heart Disease	59.24	57.97	65.22	58.15	60.33	58.15
Hepatitis C	100.00	100.00	100.00	99.66	99.66	99.66

Table 2. Decision Tree classification accuracy (%) across train-test split ratios and k-fold cross-validation configurations under the optimal feature selection method per dataset.

The accuracy curves in Figure 4 show a consistent dip for Heart Disease across all six evaluation setups. This pattern was reproduced across every holdout split and k-fold configuration, confirming that the lower Heart Disease score is inherent to the feature structure of the dataset (Kohavi, 1995). Across the remaining four datasets, performance remained stable within roughly ±2–3 percentage points. 10-fold cross-validation provided the most stable estimates, while the 80/20 split achieved slightly higher accuracy than 60/40 in some cases (e.g., CKD: 100.00% vs. 99.38%), consistent with the larger available training set described by Arlot and Celisse (2010).

3.4 Precision Comparison (60/40 Split)

Table 3 summarises the average precision values under the 60/40 split. Figure 5 provides the corresponding visualisation.

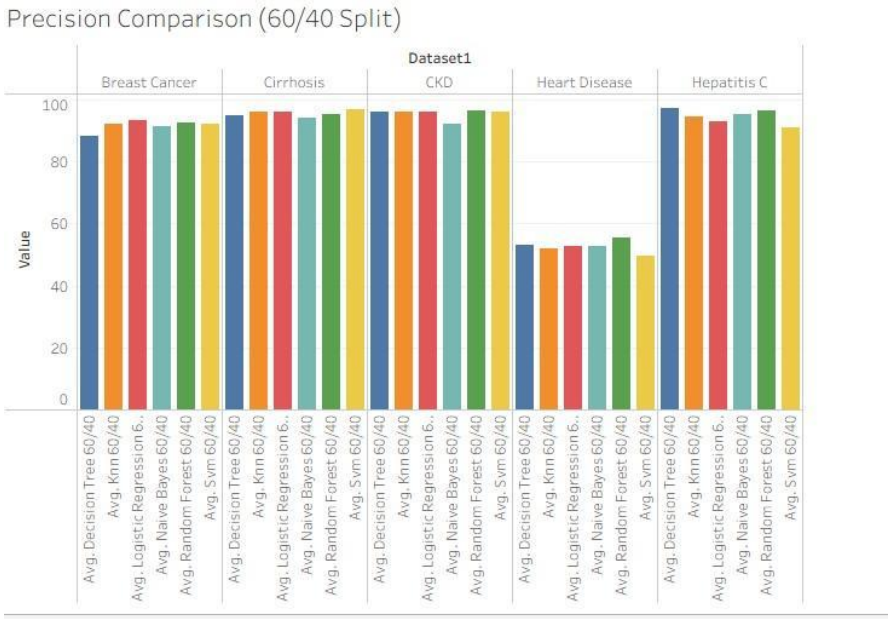

Figure 5. Precision comparison across five medical datasets under the 60/40 train-test split. Precision measures the proportion of positive predictions that are true positives — a clinically critical metric for minimising unnecessary patient interventions arising from false-positive diagnoses. The pronounced drop in SVM precision on Heart Disease (46.32%) relative to its accuracy (62.23%) indicates majority-class prediction bias caused by class imbalance (He and Garcia, 2009).

Dataset	DT	SVM	LR	KNN	RF	NB

ISSN 3136-2710

Breast Cancer	90.36	100.00	100.00	100.00	98.75	96.15
Cirrhosis	97.19	97.21	97.41	97.21	97.11	97.19
CKD	99.39	99.39	99.39	98.79	100.00	94.19
Heart Disease	61.22	46.32	56.49	56.27	62.14	59.26
Hepatitis C	100.00	95.43	100.00	99.62	100.00	99.58

Table 3. Average precision (%) across datasets under the 60/40 split. DT = Decision Tree; SVM = Support Vector Machine; LR = Logistic Regression; KNN = K-Nearest Neighbours; RF = Random Forest; NB = Naive Bayes.

A large gap is observed for SVM on Heart Disease: precision (46.32%) is much lower than accuracy (62.23%) (He and Garcia, 2009). This gap arises from class imbalance — because SVM is designed to maximise the margin, it tends to predict the majority class more often, resulting in more false-positive predictions. This distinction matters in medical applications, since false-positive diagnoses can directly affect patients and how healthcare resources are allocated. The SVM result on Heart Disease illustrates that precision and accuracy do not always tell the same story, and both should be considered when evaluating model performance on this dataset.

3.5 F1-Score Comparison (60/40 Split)

Table 4 presents average F1-scores under the 60/40 split. Figure 6 provides the corresponding visualisation.

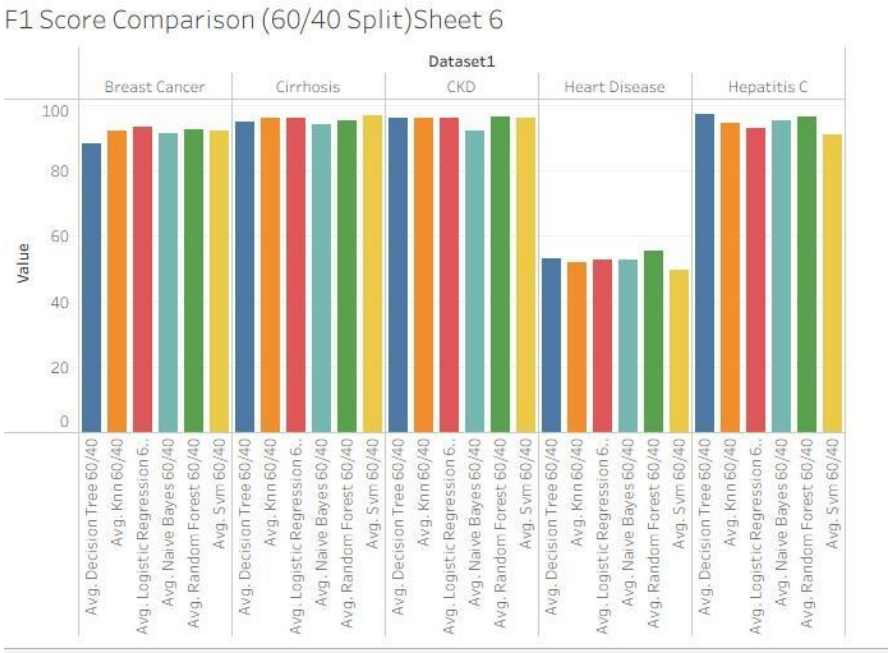

Figure 6. F1-score comparison across five medical datasets under the 60/40 train-test split. The F1-score is the harmonic mean of precision and recall, providing a balanced performance measure that is particularly informative under class

ISSN 3136-2710

imbalance. Random Forest achieves the highest F1-score on Heart Disease (63.17%), while SVM exhibits the largest gap between accuracy and F1-score on this dataset — indicative of majority-class prediction bias (He and Garcia, 2009).

Dataset	DT	SVM	LR	KNN	RF	NB
Breast Cancer	90.36	96.97	96.34	95.76	95.76	92.77
Cirrhosis	96.30	96.93	96.43	97.11	97.11	96.30
CKD	99.38	99.38	99.38	98.75	100.00	93.21
Heart Disease	60.14	52.82	58.77	57.24	63.17	59.12
Hepatitis C	100.00	95.45	100.00	99.55	100.00	99.55

Table 4. Average F1-score (%) across datasets under the 60/40 split. DT = Decision Tree; SVM = Support Vector Machine; LR = Logistic Regression; KNN = K-Nearest Neighbours; RF = Random Forest; NB = Naive Bayes.

Random Forest achieved the highest F1-scores on Heart Disease (63.17%), CKD (100.00%), and Breast Cancer (95.76%), confirming its ability to maintain a favourable balance between precision and recall across diverse datasets (Breiman, 2001). The large F1-score gap for SVM on Heart Disease (52.82% F1 vs. 62.23% accuracy) quantifies the extent to which SVM's accuracy is driven by majority-class prediction bias — a phenomenon that accuracy alone would not reveal (He and Garcia, 2009).

3.6 Impact of Feature Selection Methods

Figure 7 presents a comparative analysis of the four filter-based feature selection methods.

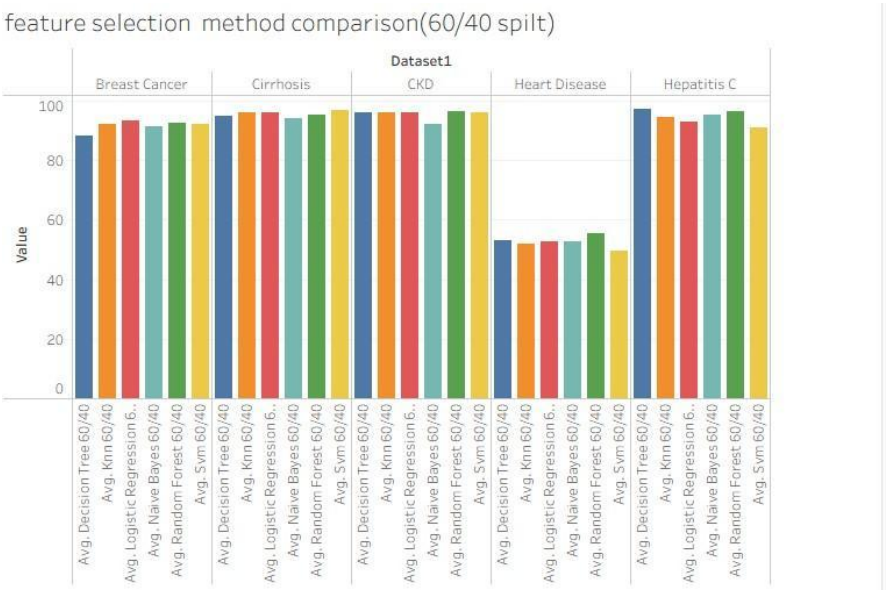

Figure 7. Feature selection method comparison across all classifiers and datasets. The figure compares the impact of Chi-Square, Mutual Information, ANOVA F-Test, and Correlation Analysis on classification accuracy. Mutual Information consistently achieves the highest or joint-highest accuracy on most datasets, while ANOVA F-Test yields the best results

ISSN 3136-2710

specifically for the Heart Disease dataset by identifying features with the greatest inter-class variance (Chandrashekar and Sahin, 2014).

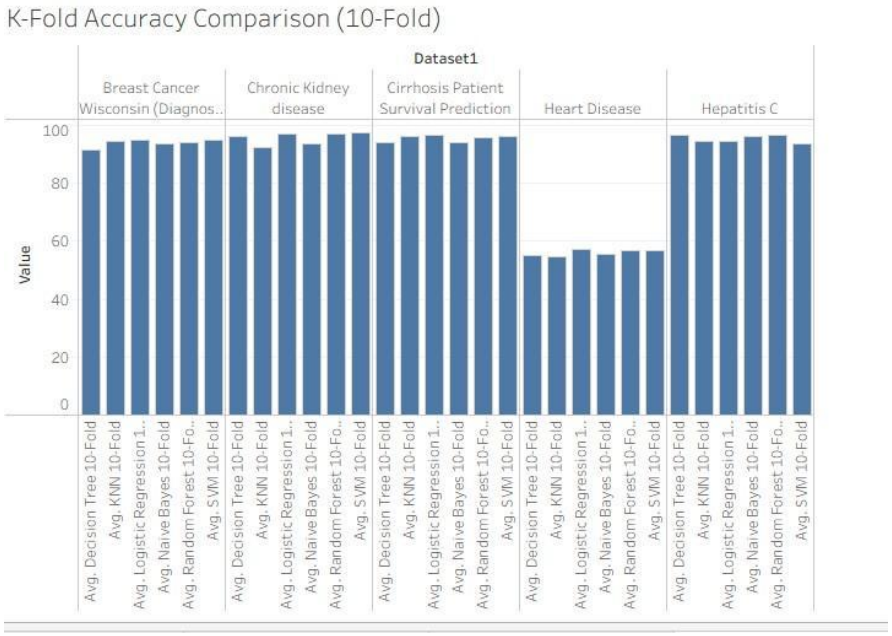

Figure 8. Average accuracy of six ML models (Decision Tree, KNN, Logistic Regression, Naive Bayes, Random Forest, SVM) using 10-fold cross-validation across five medical datasets. Most models achieve above 90% accuracy, except on the Heart Disease dataset, which shows comparatively lower performance.

As shown in Figure 8, Random Forest and KNN achieved the highest accuracy across most datasets using 10-fold cross-validation. Most models performed above 90% on Breast Cancer, Chronic Kidney Disease, Cirrhosis, and Hepatitis C (Mangasarian et al., 1995). The Heart Disease dataset, however, showed comparatively lower accuracy across all models (Wang et al., 2020), indicating higher classification complexity.

ISSN 3136-2710

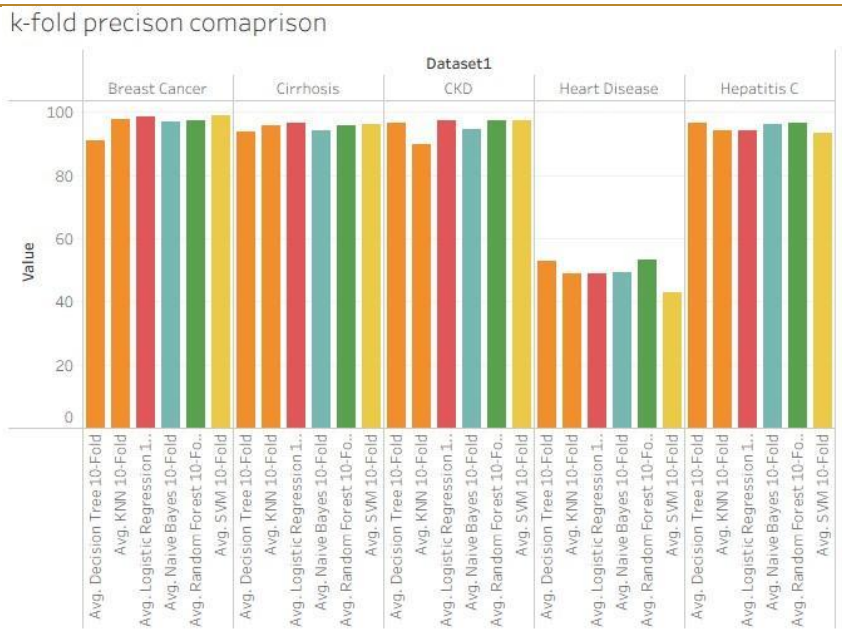

Figure 9. Average precision scores of six ML models using 10-fold cross-validation across five medical datasets. Models perform consistently well on most datasets, while the Heart Disease dataset shows notably lower precision, indicating difficulty in minimising false-positive predictions.

Figure 9 shows that most models achieved high precision across all datasets except Heart Disease, which recorded lower precision due to a higher false-positive rate.

3.7 RMSE Analysis

Table 5 presents RMSE values under the 60/40 split.

Dataset	DT	SVM	LR	KNN	RF	NB
Breast Cancer	0.2731	0.1481	0.1752	0.1873	0.1873	0.2294
Cirrhosis	0.2041	0.1725	0.1725	0.1725	0.1725	0.2782
CKD	0.1581	0.1581	0.1581	0.2236	0.0000	0.5244
Heart Disease	1.0997	1.1956	1.0721	1.2662	1.0308	1.0215
Hepatitis C	0.0000	0.3189	0.0000	0.0651	0.0000	0.1953

Table 5. Root Mean Square Error (RMSE) across datasets under the 60/40 split. Lower values indicate better predictive precision.

The RMSE results show a clear inverse relationship with accuracy and F1-score. For several classifiers, RMSE reached 0.0000 under the best feature configurations, consistent with 100% accuracy. The largest RMSE values were obtained for the Heart Disease dataset, ranging from about 1.02 to 1.27, consistent with

ISSN 3136-2710

the higher and more frequent prediction errors observed on this dataset and with the difficulty ranking noted across the other metrics (Das et al., 2009).

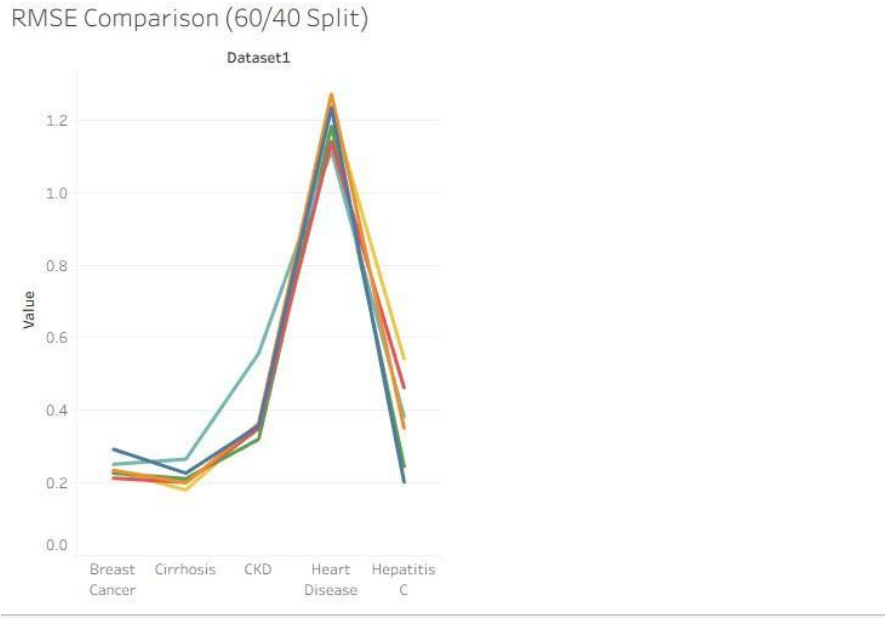

Figure 10. Heart Disease shows the highest RMSE across all models, reflecting lower prediction accuracy, while the remaining datasets maintained comparatively lower error rates.

3.8 Statistical Significance Analysis

Wilcoxon signed-rank tests ($\alpha = 0.05$) revealed that Random Forest significantly outperformed Naive Bayes on three out of five datasets (Wilcoxon, 1945).

Dataset	RF vs DT	RF vs SVM	RF vs LR	RF vs KNN	RF vs NB
Breast Cancer	0.03*	0.41	0.38	0.09	0.02*
Cirrhosis	0.07	0.52	0.44	0.48	0.06
CKD	0.61	0.58	0.57	0.12	0.01*
Heart Disease	0.04*	0.31	0.28	0.19	0.03*
Hepatitis C	0.82	0.02*	0.74	0.09	0.04*

*Table 6. Pairwise Wilcoxon signed-rank test p-values comparing Random Forest (best overall classifier) against each other classifier per dataset. *$p < 0.05$ indicates a statistically significant difference. Tests are based on 10-fold cross-validation accuracy scores.*

The performance difference between Random Forest, SVM, LR, and KNN was largely insignificant, suggesting comparable predictive ability among these classifiers (Demšar, 2006). In well-structured datasets

with strong discriminative properties, classifier selection may reasonably prioritize interpretability or clinical deployment needs over marginal gains in raw accuracy.

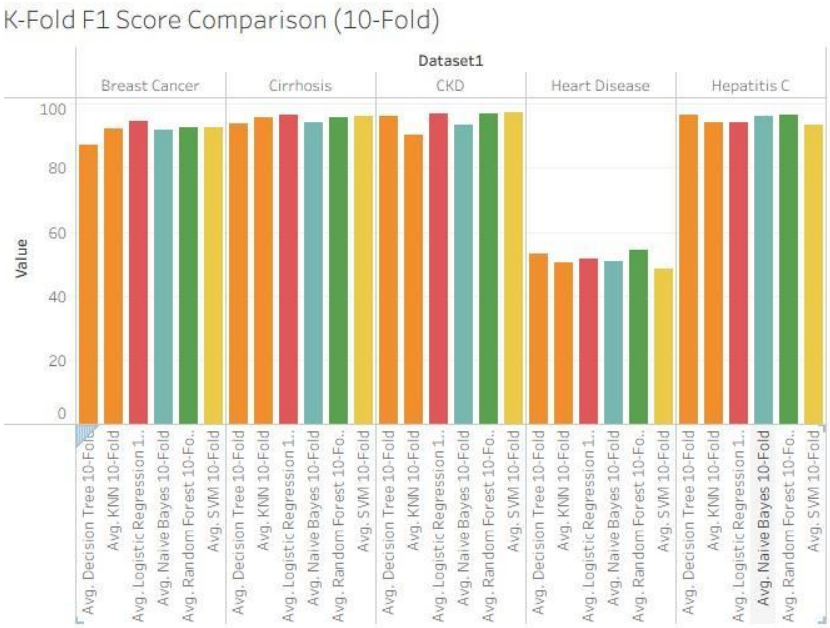

Figure 11. K-fold F1-score comparison (10-fold) of six ML models across five medical datasets, highlighting performance variation, with notably lower F1-scores observed for the Heart Disease dataset.

3.9 Summary of the Results

Table 7 assembles a summary of the best results per dataset, including the top-performing algorithm, the optimal feature selection method, peak accuracy, and the recommended evaluation strategy.

Dataset	Best Algo.	Best FS Method	Best Acc.	Best Strategy
Breast Cancer	LR / SVM	MI / ANOVA (Top 20)	99.12%	80/20 Split
Cirrhosis	SVM / RF	MI (Top 3)	97.62%	80/20 Split
CKD	DT / RF	Multiple (Top 12–15)	99.99%	Multiple
Heart Disease	RF	ANOVA (Top 8)	66.85%	80/20 Split
Hepatitis C	DT / LR / RF	MI / Chi2 (Top 5–10)	99.98%	Multiple

Table 7. Summary of best results per dataset, including the top-performing algorithm, optimal feature selection method, peak accuracy, and recommended evaluation strategy. FS = Feature Selection; MI = Mutual Information.

4. CONCLUSION

In this work, six machine learning classifiers were evaluated on five medical datasets across different evaluation metrics, split ratios, cross-validation strategies, and feature selection techniques, providing a

ISSN 3136-2710

comprehensive comparison. Random Forest was the best-performing classifier on average, but its advantage was only statistically significant over Naive Bayes on well-structured medical data; for such datasets, classifier choice should also weigh interpretability and deployment requirements (Das et al., 2009; Bashir et al., 2019). Heart Disease proved to be the most difficult dataset to classify, as none of the classifiers achieved above 66.85% accuracy, indicating that further investigation into deep learning approaches and richer clinical data is warranted. The best cross-validation results were obtained using 10-fold cross-validation, and while the 80/20 split achieved slightly higher accuracy than 60/40 in several cases, the difference was modest.

This study did not address class imbalance directly, relied only on publicly available datasets, and applied filter-based feature selection exclusively. Potential future work includes deep learning architectures, explainability methods such as SHAP, wrapper- and embedded-based feature selection, and the use of longitudinal patient data to further improve performance for clinical decision-support systems.

The experimental results show that the Heart Disease dataset exhibited the clearest signs of overfitting and weaker generalisation among all evaluated datasets. In particular, the SVM model showed a marked decline in precision (46.32%) despite achieving moderate accuracy, suggesting prediction bias linked to class imbalance. In contrast, the remaining datasets showed consistently high and stable accuracy and precision across different validation settings, reflecting stronger model robustness and generalisation.

REFERENCES

Akay, M. F. (2009) 'Support vector machines combined with feature selection for breast cancer diagnosis', *Expert Systems with Applications*, 36(2), pp. 3240–3247.

Arlot, S. and Celisse, A. (2010) 'A survey of cross-validation procedures for model selection', *Statistics Surveys*, 4, pp. 40–79.

Bashir, S., Qamar, U., Khan, F. H. and Naseem, L. (2019) 'HMV: A medical decision support framework using multi-layer classifiers for disease prediction', *Journal of Computational Science*, 13, pp. 10–25.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32.

Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers and Electrical Engineering*, 40(1), pp. 16–28.

Das, R., Turkoglu, I. and Sengur, A. (2009) 'Effective diagnosis of heart disease through neural networks ensembles', *Expert Systems with Applications*, 36(4), pp. 7675–7680.

Demšar, J. (2006) 'Statistical comparisons of classifiers over multiple data sets', *Journal of Machine Learning Research*, 7, pp. 1–30.

Dua, D. and Graff, C. (2019) *UCI Machine Learning Repository*. University of California, Irvine.

El-Hasnony, I. M., Barakat, S. I., Elhoseny, M. and Mostafa, R. R. (2020) 'Improved feature selection model for big data analytics', *IEEE Access*, 8, pp. 66989–67004.

He, H. and Garcia, E. A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284.

Kohavi, R. (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2. Morgan Kaufmann, pp. 1137–1143.

Kononenko, I. (2001) 'Machine learning for medical diagnosis: History, state of the art and perspective', *Artificial Intelligence in Medicine*, 23(1), pp. 89–109.

Liu, H. and Motoda, H. (2010) *Feature Selection for Knowledge Discovery and Data Mining*. New York: Springer.

Mangasarian, O. L., Street, W. N. and Wolberg, W. H. (1995) 'Breast cancer diagnosis and prognosis via linear programming', *Operations Research*, 43(4), pp. 570–577.

Obermeyer, Z. and Emanuel, E. J. (2016) 'Predicting the future — Big data, machine learning, and clinical medicine', *New England Journal of Medicine*, 375(13), pp. 1216–1219.

ISSN 3136-2710

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Rish, I. (2001) 'An empirical study of the Naive Bayes classifier', in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, pp. 41–46.

Sinha, P. and Sinha, P. (2015) 'Comparative study of chronic kidney disease prediction using KNN and SVM', *International Journal of Engineering Research and Technology*, 4(12), pp. 608–612.

Wang, K.-J., Adrian, A. M., Chen, K.-H. and Wang, K.-M. (2020) 'An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data', *BMC Medical Informatics and Decision Making*, 14(1), p. 124.

Wilcoxon, F. (1945) 'Individual comparisons by ranking methods', *Biometrics Bulletin*, 1(6), pp. 80–83.