*International Journal of Basic and Applied Sciences*, 14 (4) (2025) 269-278

**International Journal of Basic and Applied Sciences**

*Website: www.sciencepubco.com/index.php/IJBAS*
*https://doi.org/10.14419/xn4vbz19*
***Research paper***

# Predictive Modelling of cardiovascular Disease Survival Using Mutual Information and Machine Learning Across Varying Sample Sizes

**Vijayalakshmi. Sarraju [1] *, Jaya pal [1], Supreeti. Kamilya [2]**

[1] *Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Lalpur, Ranchi, India*
[2] *Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India*
**Corresponding author E-mail: vijayalakshmi@bitmesra.ac.in*

*Received: June 27, 2025, Accepted: August 3, 2025, Published: August 11, 2025*

**Abstract**

In clinical data analytics, predicting survival outcomes for cardiovascular disease (CVD) is a challenging task with practical implications. Using three different datasets, this study investigates how sample size affects machine learning performance and generalizability. The methodology combines statistical sample-size analysis with mutual information gain, a filter-based, scalable, and domain-agnostic feature selection strategy, to identify clinically essential features. Mutual information gain measures the dependency between each predictor and the target variable, ensuring computational efficiency and applicability to large-scale data. Machine learning classifiers, support vector machines (SVMs) and logistic regression (LR), are employed to assess predictive performance across varying population sizes. Experimental results demonstrate that increasing the sample size improves model accuracy by up to 10%, recall by 5–8%, and maintains consistent specificity. Furthermore, to enhance clinical reliability, the models are evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), where SVM achieved an AUC of 0.965 and LR achieved 0.937, validating strong discriminatory power; Also, SHAP-based feature attribution is used to improve interpretability, identifying that larger sample sizes provide more stable and clinically meaningful explanations.

***Keywords***: *Cardiovascular Disease (CVD); Support Vector Machine (SVM); Logistic Regression (LR); Sample Size.*

## 1. Introduction

Health professionals diagnose cardiovascular disease (CVD) using complex clinical and pathological data, often employing numerous diagnostic tests, specialist opinions, and prolonged medical assessments [1]. Because of a more extended diagnosis period and more resources, the complexity of cardiovascular illness results in additional expenses in the provision of medical treatment. Furthermore, reducing treatment appropriateness and minimising patient care quality are complications in detecting essential risk factors and a lack of consistency in diagnostic techniques.

The World Health Organisation indicates that cardiovascular disease affects one-third of individuals in underdeveloped nations. American Heart Association: One-third of people suffer from cardiovascular illnesses. Predictive models can improve diagnosis by using diverse data combinations and expert insights. The prediction process requires numerous statistical studies and machine learning algorithms. Identifying hidden medical information in clinical data from various health symptoms and individuals with cardiovascular disease (CVD) is an essential and practical approach for diagnosing CVD by employing clinical data, statistics, and machine learning (ML) to forecast stages of heart disease. Using multiple data configurations and expert knowledge, prediction models can enhance diagnosis. This prediction process involves various statistical analyses and machine learning models [2-3]. Finding concealed medical information in clinical data from diverse manifestations of health and individuals with CVD is a distinguished and effective strategy for classifying cardiovascular disease and using clinical data, statistics, and machine learning (ML) to predict heart disease stages [4-5].

Machine learning algorithms can predict cardiovascular disease by analysing complex patterns and risk factors in large datasets. These technologies can quickly identify high-risk patients for personalised treatment. Medical history, genetics, lifestyle, and biomarkers are used to develop accurate prediction models. These algorithms help clinicians determine ongoing treatments and appropriate patient care to reduce cardiovascular disease mortality. Research will show how SVM and LR can predict cardiovascular disease. Each technique has its advantages and disadvantages depending on the specific investigation's objectives. Logistic regression considers age, cholesterol, and blood pressure as CVD risk factors. Individual risk factors influence patient treatment and clinical decision-making. SVM kernels capture nonlinear interactions among variables and efficiently manage multidimensional datasets [6-9].

A statistical method, mutual information gain, is employed to identify the most significant clinical characteristics for model training. These strategies enhance model efficacy and reduce computing costs by identifying variables that are significantly correlated with a specific outcome.

The study proposes improving cardiovascular disease prediction through simple random sampling, statistical feature selection, and SVM and Logistic Regression techniques. The investigation utilised the Cleveland dataset, a comprehensive dataset of Statlog, Cleveland, Switzerland, Hungary, and Long Beach, VA, along with a dataset focused on stroke prediction. The investigation was conducted thoroughly, including a comprehensive sampling of datasets of varying sizes to ensure statistical representation. Inferential statistics are employed to determine sample sizes and ensure representativeness. Mutual information gain is utilised to identify the risk of cardiovascular disease. Logistic Regression and Support Vector Machine classifiers were trained and tested based on defined characteristics. The characteristics are utilised to train and evaluate LR and SVM classifiers. The accuracy of the classifier is enhanced as the sample sizes increase, achieving rates between 80% and 95% during 5-fold and 10-fold cross-validation. The findings indicate that SVM better identifies CVD risk factors than LR. This investigation analyses sample size, feature selection, and the performance of classifiers for CVD prediction, emphasising the importance of statistical inference in clinical decision-making.

The structure of the research article. Section 2 provides an overview of the statistical and machine learning literature on CVD. Section 3 discusses the dataset description and methodology. Section 4 presents an analysis of the results. Section 5 represents the discussion. The article concludes in Section 6.

## 2. Literature review

This survey presents well-established statistical learning-based cardiovascular disease diagnostic approaches to demonstrate the relevance of the proposed study.

Using principal component analysis and feature selection techniques, Santhanam and Ephzibah (2013) created a prognostic model using regression and feed-forward neural networks, achieving 95.2% accuracy in the Cleveland heart disease dataset and six additional datasets [10]. Ziasabounchi and Askerzade (2014) used PCA to extract characteristics for clustering, resulting in improved accuracy from 81.0% to 87.0% and 82.0%, respectively [11]. A univariate feature evaluator and a random forest were used to evaluate Statlog heart disease data by Jabbar et al (2015). After backward elimination, the model was 83.7% accurate [12]. To minimise data size, Kavitha and Kannan (2016) recommended PCA for feature extraction. This decreased processing resources and enhanced model accuracy [13].

Khateeb and Usman (2017) evaluated NB, KNN, decision tree, and bagging methods on the Cleveland dataset. By selecting context-relevant variables, they improved model accuracy, achieving 92% accuracy with KNN [14]. Gokulnat and Shantharajah used a genetic algorithm and four machine-learning algorithms to select features. SVM demonstrated the highest accuracy of 88.34%, significantly exceeding the original dataset's 83.70% [15]. In (2020) [16], Mehmet S´ata and others studied how sample size influences categorization using LR and CHAID. This research employed the "Attentional Control Scale." The logistic regression classifier and CHAID tests were evaluated and interpreted using nine parameters. Logistic regression outperformed CHAID analysis with small sample sizes (N=25–900) and remained consistent. CHAID classification improved with N=1000+ samples.

In 2021, Harshit Jindal et. al [17] determined whether medical factors enhance heart disease risk. Medical history informs their heart disease prediction system. KNN and logistic regression classified cardiac patients. NB KNN and LR correctly predicted cardiac disease. In 2022, Chaimaa Boukhatem et.al [18] explained how machine-learning methods may predict heart illness using patient health data. The article used MLP, SVM, RF, and NB classification algorithms to create prediction models. Model construction was preceded by feature selection and data preparation. Model performance was measured. The SVM model has the highest accuracy (91.67%). In 2023 [19], A study by Kavya S M et al. used Kaggle dataset heart disease prediction models to estimate 10-year CHD risk using patient cardiac parameters and LR. The collection includes 4,000 records, 15 risk factors and patient data. It classifies binary and multi-class. LR calculates heart disease risk from factors. To enhance logistic regression (ILR) using multiple classification methodologies, Arkadip, Ray, et al. advocated ILR classifier assessment in 2024 [20]. RF, SVM, NB, and DT are major machine learning classifiers. Compare the suggested strategy. The ILR model's 83% classification rate is favourable.

Mityoshi Takara and others investigated baseline variables and CVD risk in Japanese diabetics aged 40–74[21]. A prospective multicentre cohort of 6338 (634 with CVD and 5704 without) was studied. After adjusting for non-cardiovascular mortality, a competing risk model showed 413 CVD occurrences with 8-year cumulative incidence rates of 21.5% in CVD patients and 7.2% in non-CVD patients over 6.9 years. Those without a history of CVD had higher systolic blood pressure and lower HDL cholesterol, with interaction P values of 0.040 and 0.005, respectively. Male sex, older age, longer diabetes duration, higher haemoglobin A1c, and higher LDL cholesterol were risk factors regardless of CVD history. The study demonstrated that metabolic profiles predict CVD risk differently in those with and without a CVD history, suggesting that primary prevention risk calculators may be ineffective for secondary prevention in this population.

### 2.1. Research gap

Previous studies on cardiovascular disease (CVD) prediction have extensively used machine -learning models and feature extraction techniques such as PCA, genetic algorithms, and univariate selection. However, few works have explored filter-based feature selection methods, such as mutual information gain, combined with interpretable classifiers like logistic regression and SVM, to identify clinically significant predictors statistically comprehensively. Sample-size-driven studies report model accuracy on fixed datasets; the effect of varying sample sizes on model performance, particularly generalisation and stability, has not been systematically investigated. Most models are evaluated without validating how performance metrics change as data volume increases, which is critical in real-world clinical applications with imbalanced datasets. Additionally, many existing works focus on scalar metrics like accuracy or F1-score but lack statistical validation or visual performance diagnostics (e.g., ROC curves, calibration plots) to assess model reliability. These limitations create a gap in developing robust, scalable, and interpretable diagnostic models. This study addresses the gaps mentioned in the research by introducing a sample–size–driven framework that uses mutual information gain for feature selection and compares the predictive efficiency of SVM and logistic regression across multiple datasets with detailed statistical evaluation.

### 2.2. Research framework

Figure 1: The workflow diagram of the proposed model shows that the pipeline begins with input from raw clinical data, followed by preprocessing steps like missing value imputation and normalisation. A statistical feature selection process using Mutual Information.

Selected features are then used to train classifiers such as Logistic Regression (LR) and Support Vector Machine (SVM). The model is evaluated using AUC-ROC, accuracy, and F1-score to ensure clinical reliability and robustness.

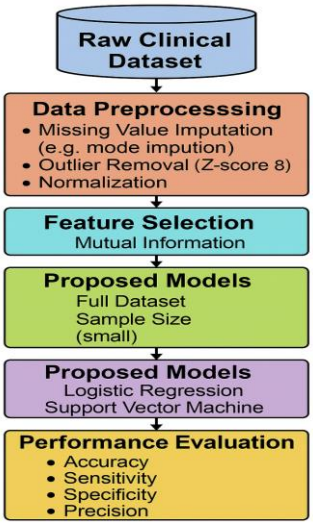


**Fig. 1:** Workflow Diagram of Clinical Data-Based Machine Learning Pipeline.

# 3. Methodology

## 3.1. Dataset

This study uses three medical datasets to evaluate predictive modelling for cardiovascular and stroke risk: (i) the Cleveland Heart Disease dataset (UCI repository),[22] (ii) a comprehensive heart disease dataset compiled from Cleveland, Statlog, Hungary, Switzerland, and Long Beach VA sources,[23] and (iii) a stroke prediction dataset from Kaggle [24]. The Cleveland dataset contains 303 records and 13 clinical features, including age, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, ECG results, maximum heart rate, and exercise-induced angina, with the target indicating the presence of heart disease. The comprehensive dataset combines similar features from five data sources, creating a feature-rich, heterogeneous dataset with 1190 instances and 12 attributes. The stroke prediction dataset has 5110 entries and includes 11 features that combine demographic and health indicators, such as gender, age, hypertension, heart disease status, BMI, average glucose level, smoking status, and the binary target variable indicating stroke occurrence.

## 3.2. Data preprocessing

Every dataset undergoes cleaning procedures during the preprocessing phase to ensure data integrity. For outlier detection, we use the Z-score method on each numerical attribute within all three datasets—Cleveland, Comprehensive, and Stroke. Data points with a Z-score greater than ±3 are identified as outliers and removed to prevent skewing the analysis. Min-max normalisation is applied explicitly to the comprehensive dataset because it contains features with negative values, which can affect distance-based learning algorithms. This scaling transforms each feature value to a [0, 1] range, allowing features to contribute equally during training. The stroke and Cleveland datasets do not require Min-Max scaling, as their features are already within comparable ranges. These preprocessing steps ensure that the machine learning models receive clean, consistent inputs across all datasets, enhancing training stability and predictive accuracy.
The Cleveland dataset had 302 out of 303 cases, the comprehensive dataset 918 out of 1190, and the stroke prediction dataset 4228 out of 5110 after eliminating duplicates. Outliers were detected using Z-scores. The empirical approach identifies outliers when the z-score is greater than 3. However, comprehensive dataset values were normalised using Min-Max scaling since they have negative attribute column values. The values in the min-max scaler range from 0 to 1. Equation 1 refers to the min-max scalar. The z-score normalisation is calculated by scaling the distance of a data point (x) from the mean (μ) by the standard deviation (σ), referring to the z-score equation 2.

$$X_{scaled} = \frac{(X - X\_min)}{X\_max - X\_min} \quad (1)$$

$$Z = \frac{(\ddot{x} - \mu)}{\sigma} \quad (2)$$

## 3.3 Sampling approach

The sample size for the proposed dataset is determined based on the Z-score at 95-99% confidence intervals, a 50% population proportion, and a 5% margin of error. The parametric test (t-test) evaluates hypotheses to see if the sample mean significantly differs from the population mean. The null hypothesis assumes no significant difference exists. The suitability of the calculated sample size for the dataset is confirmed by accepting the null hypothesis. From this experiment, it can be concluded that the computed sample reliably represents the entire population. The sample sizes for the datasets are as follows: 170 and 280 instances for the Cleveland dataset; 291 and 428 instances for the comprehensive dataset; and 358 and 592 instances for the stroke prediction dataset. Equation 3 illustrates testing whether the sample mean ($\bar{x}$) significantly differs from the known population mean (μ), considering the sample size (n) and sample standard deviation (s).

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (3)$$

### 3.4. Predictor selection approaches

Predictor selection, a vital task in data science, involves identifying the most relevant attributes of the original characteristics that significantly influence the outcome. The advantages of attribute selection include enhanced data quality, reduced computational time for prediction models, and improved predictive accuracy. Filter-based feature selection methods use statistical techniques that analyse similarity, dependency, information, and distance to determine input-target relationships. This study employed the powerful Mutual Information Gain test to ensure the most accurate results in identifying significant features from data sets.

#### 3.4.1. Mutual information gain

The statistical filter method of mutual information gain determines the correlation between variables. Choose the aspects related to the target variable that are the most informative. Equation 4 describes the information that can be gained from two variables, X and Y, given an n-th observation. We denote the entropy of Y as H(Y), which measures the uncertainty present in a discrete random variable. H (Y | X) represents the conditional entropy of X given Y.

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{4}$$

The procedure for applying mutual information gain to the datasets is as follows:

- Choose the features of the proposed dataset.

The mutual info class if() function analyses the mutual information scores to determine the association between the dependent and independent variables—the successive formulas implemented in computing mutual information scores (Eq. 5).

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{5}$$

p(x,y) is the probability mass function. A higher value of mutual information indicates a strong dependence between the two discrete random variables, X and Y. Statistical independence between the variables suggests a mutual information gain value of zero

- A higher score signifies a greater dependence on the output variable. The SelectKBest() function is applied to identify the attributes with the highest mutual information score for constructing the model.

### 3.5. Model training and testing

The Cleveland training dataset has 14 attributes, while the comprehensive dataset has 12 attributes, and the stroke prediction dataset has 12 attributes. Performance metrics are used to validate the SVM and logistic regression models. The data is split using the train-test validation technique. For testing, the ideal split is 20%, and for training, 80%. This method can lead to overfitting because the evaluation depends on dividing the data into training and testing sets. As a result, the outcomes may vary significantly depending on the split. Cross-validation is introduced as a more reliable method to prevent overfitting. In subsequent experiments, both 5-fold and 10-fold cross-validation are used. The original dataset is divided into five equal parts, ensuring each part is used once for testing and multiple times for training. This approach improves the reliability of the evaluation by making better use of the data for both training and testing.

### 3.6. Model interpretability using SHAP

To enhance the interpretability of our prediction results, we employed SHAP (Shapley Additive Explanations), an integrated approach to explain the output of any machine learning model. SHAP assigns each feature an importance value for a particular prediction, enabling both global and local interpretability. In this study, we applied SHAP analysis across different sample sizes to observe how feature contributions vary with the quantity of training data. This enables us to assess whether feature impact remains consistent as sample size increases, suggesting insights into the stability of model interpretation.

## 4. Results and discussion

### 4.1. Results of feature selection methods

The following graphs visualise the critical characteristics identified using the mutual information gain approach and display the attributes chosen by three proposed datasets. These tables include the feature scores visualised in Fig. 2, Fig.3, and Fig.4. The features are ranked in descending order depending on their significance as evaluated by the feature selection procedure.

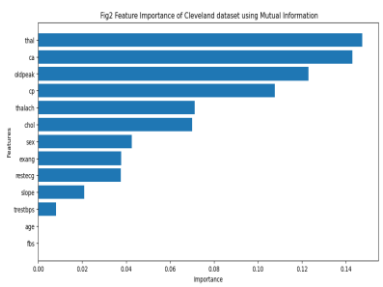


**Fig. 2**
Feature importance of the Cleveland data set.

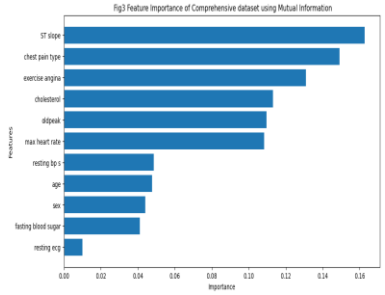


**Fig. 3**
Feature importance of the Comprehensive dataset.

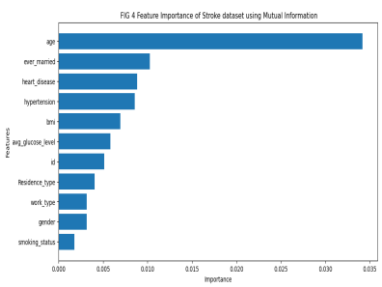


**Fig.4**
Feature importance of the Stroke dataset

#### 4.1.1. Statistical analysis and interpretation

Our work has produced Mutual Information feature significance charts for the Stroke, Comprehensive, and Cleveland datasets that show significant predictors in each category. The attributes 'thal' (outcome of the Thallium stress test), 'ca' (number of major vessels visualised via fluoroscopy), and 'old peak' (exercise-induced ST segment depression compared to rest) in the Cleveland dataset are notable, underscoring the physiological responses to stress and the condition of the heart's vessels as critical indicators in predicting heart disease. The Comprehensive dataset underscores 'ST slope' (the peak exercise ST segment slope), 'chest pain type', and 'exercise angina' as primary variables, accentuating the severity of activity-related cardiac symptoms and their manifestation patterns in diagnosing coronary artery disease. In the Stroke dataset, 'age', 'job type', and 'BMI' (Body Mass Index) are paramount, highlighting the substantial influence of biological age and lifestyle determinants, including body weight and profession, on stroke risk. chronic findings highlight the multifaceted cause of chronic health disorders, wherein lifestyle and physiological variables are crucial. These discoveries may substantially improve patient care by facilitating more focused screening and patient enquiries in clinical settings.

### 4.2. Performance evaluation of classifiers on three datasets

Heart disease classification uses supervised machine learning methods like SVM and logistic regression, ensuring accuracy with five- and ten-fold cross-validation. Samples were classified as non-disease (unfavourable) or disease (positive). Performance metrics such as true positives, false positives, and false negatives assess model dependability. Classification accuracy replicates the percentage of appropriately projected samples. Recall (sensitivity) quantifies true positives among total positives, while specificity measures the accurate identification of negatives. Positive predictive value evaluates true positives within predicted positives. The F1 score, a robust metric that balances precision and recall, ensures the model's performance, with 0 and 1. Equations 6, 7, 8, 9, and 10 show the calculations of accuracy, sensitivity, specificity, precision, and F1 score.

$$\text{Accuracy (Acc)} = \frac{TP+TN}{FP+FN+TP+TN} \tag{6}$$

$$\text{Sensitivity (Sen)} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Specificity(spe)} = \frac{TN}{TN+FP} \tag{8}$$

$$\text{Precision (Prec)} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{F1score(F1-Sco)} = 2\frac{\text{Precision}*\text{sensitivity}}{\text{precision}+\text{sensitivity}} \tag{10}$$

We can evaluate the prediction model's performance and validity by generating these measures.

#### 4.2.1. Analysis of classifier performance on the selected feature set for the Cleveland dataset

Tables 1 and 2 show the performance evaluation of the SVM and LR classifiers on the original Cleveland dataset and subsets with sample sizes of 170 and 208, which are considered representative of medium and large datasets commonly encountered in healthcare analytics. The analysis of SVM and LR classifiers on the Cleveland dataset for cardiovascular disease prediction reveals distinguished differences in their performance metrics. SVM consistently demonstrates higher accuracy, specificity, sensitivity, precision, and F1-score than LR across various scenarios and sample sizes, underscoring its reliability. Both 5-fold and 10-fold cross-validation techniques affirm SVM's stability and reliability, maintaining accuracy above 93% with balanced sensitivity and specificity. Additionally, SVM exhibits robust performance even with smaller sample sizes, with accuracy above 90%. In contrast, LR's performance shows more variability, particularly during cross-validation, where accuracy fluctuates around 84% to 85%. LR retains competitive accuracy in bigger sample sets, with a 94% accuracy for 170 and over 92% for 208. The graphical presentation is shown in Fig. 5.

**Table 1:** Performance Metrics of SVM Classifier on Cleveland Dataset.

| Metrics | 80-20 split (Original Data) | 5-fold CV (Original data) | 10-fold CV (Original data) | 80-20 split (sample size 170) | 5-fold CV (sample size 170) | 10-fold CV (sample size 170) | 80-20 split (Sample size 208) | 5-fold CV (Sample size 208) | 10-fold CV (Sample size 208) |
|---|---|---|---|---|---|---|---|---|---|
| accuracy | 0.90 | 0.93 | 0.93 | 0.91 | 0.86 | 0.89 | 0.92 | 0.91 | 0.92 |
| specificity | 0.90 | 0.94 | 0.94 | 0.9 | 0.89 | 0.91 | 1.0 | 0.96 | 0.93 |
| sensitivity | 0.89 | 0.92 | 0.92 | 0.92 | 0.83 | 0.86 | 0.89 | 0.86 | 0.92 |
| precision | 0.90 | 0.94 | 0.94 | 0.91 | 0.88 | 0.90 | 0.94 | 0.96 | 0.93 |
| F1-score | 0.90 | 0.93 | 0.93 | 0.91 | 0.85 | 0.88 | 0.93 | 0.90 | 0.92 |

**Table 2:** Performance of LR Classifier on Selected Feature Set for the Cleveland Dataset for Heart Disease Prediction

| Metrics | 80-20 split (Original Data) | 5-fold CV (Original data) | 10-fold CV (Original data) | 80-20 split (sample size 170) | 5-fold CV (sample size 170) | 10-fold CV (sample size 170) | 80-20 split (Sample size 208) | 5-fold CV (Sample size 208) | 10-fold CV (Sample size 208) |
|---|---|---|---|---|---|---|---|---|---|
| accuracy | 0.86 | 0.84 | 0.84 | 0.94 | 0.82 | 0.83 | 0.92 | 0.89 | 0.89 |
| specific-ity | 0.88 | 0.84 | 0.84 | 0.92 | 0.81 | 0.81 | 1.000 | 0.91 | 0.91 |
| sensitiv-ity | 0.85 | 0.84 | 0.85 | 0.95 | 0.84 | 0.85 | 0.70 | 0.87 | 0.88 |
| precision | 0.87 | 0.83 | 0.85 | 0.93 | 0.81 | 0.83 | 0.94 | 0.92 | 0.91 |
| F1-score | 0.88 | 0.83 | 0.84 | 0.93 | 0.82 | 0.83 | 0.95 | 0.89 | 0.89 |

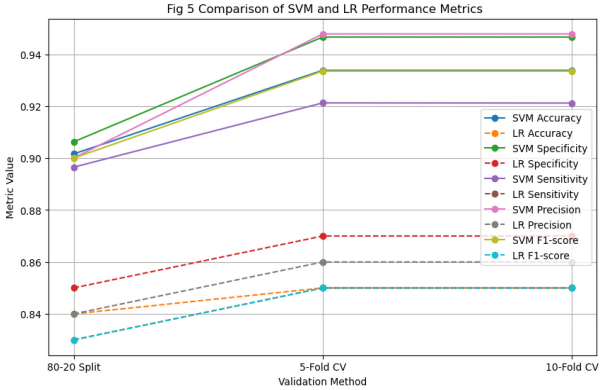


**Fig. 5:** Comparison of SVM and LR Performance Metrics on the Cleveland Dataset.

### 4.2.2. Analysis of classifier performance on combined dataset (of size 1190)

The performance of SVM and LR classifiers on the selected feature set is thoroughly analysed through statistical measures and cross-validation techniques, as shown in Tables 3 and 4. SVM initially outperforms LR in accuracy, specificity, sensitivity, precision, and F1-score on the original dataset. On the original dataset, 5-fold and 10-fold cross-validation validated SVM's stability and generalisation reliability. Even when applied to smaller sample sizes of 291 and 428 instances, SVM maintained high accuracy levels, demonstrating its robustness across different dataset sizes. In contrast, LR's performance exhibited a noticeable decline during cross-validation, with accuracy dropping to around 80%, highlighting the importance of sample selection in optimising predictive models. The graphical representation of SVM and LR is shown in Fig. 6, respectively.

**Table 3:** Performance of SVM Classifier on Selected Feature Set for Comprehensive Dataset

| Metric | 80-20 split (Original data) | 5-fold CV (Original data) | 10-fold CV (Original data) | 80-20 split (sample size 291) | 5-fold CV (sample size 291) | 10-fold CV (sample size 291) | 80-20 split (sample size 428) | 5-fold CV (sample size 428) | 10fold CV (sample size 428) |
|---|---|---|---|---|---|---|---|---|---|
| accuracy | 0.93 | 0.94 | 0.94 | 0.94 | 0.93 | 0.92 | 0.94 | 0.92 | 0.92 |
| specific-ity | 0.97 | 0.95 | 0.95 | 1.0 | 0.91 | 0.91 | 0.98 | 0.91 | 0.91 |
| sensitiv-ity | 0.92 | 0.93 | 0.93 | 0.92 | 0.94 | 0.94 | 0.89 | 0.94 | 0.94 |
| preci-sion | 0.94 | 0.95 | 0.95 | 0.96 | 0.91 | 0.91 | 0.94 | 0.91 | 0.91 |
| F1-score | 0.90 | 0.94 | 0.94 | 0.92 | 0.92 | 0.92 | 0.95 | 0.93 | 0.93 |

**Table 4:** Performance of LR Classifier on Selected Feature Set for Comprehensive Dataset

| Metrics | 80-20 split (Original data) | 5-fold CV (Original data) | 10-fold CV (Original data) | 80-20 split (sample size 291) | 5-fold CV (sample size 291) | 10-fold CV (sample size 291) | 80-20 split (sample size 428) | 5-fold CV (sample size 428) | 10fold CV (sample size 428) |
|---|---|---|---|---|---|---|---|---|---|
| accuracy | 0.93 | 0.80 | 0.80 | 0.93 | 0.82 | 0.83 | 0.92 | 0.85 | 0.84 |
| specificity | 0.94 | 0.80 | 0.80 | 0.94 | 0.82 | 0.82 | 0.90 | 0.86 | 0.86 |
| sensitivity | 0.92 | 0.82 | 0.82 | 0.92 | 0.83 | 0.83 | 0.92 | 0.83 | 0.82 |
| precision | 0.94 | 0.80 | 0.80 | 0.94 | 0.81 | 0.83 | 0.92 | 0.86 | 0.85 |
| F1-score | 0.90 | 0.81 | 0.81 | 0.90 | 0.82 | 0.83 | 0.92 | 0.84 | 0.83 |

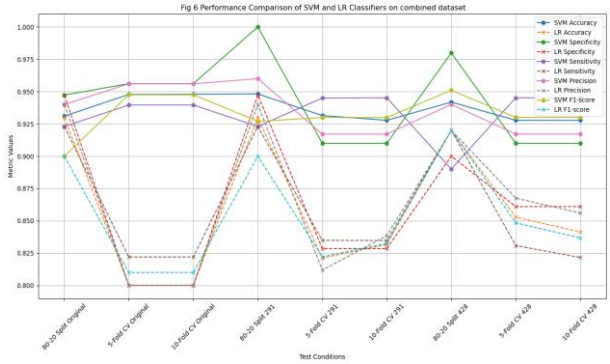


**Fig. 6:** Comparison of SVM and LR Performance Metrics on the Comprehensive Dataset.

### 4.2.3. Analysis of classifier performance on the stroke prediction dataset (of size 5110)

Evaluating SVM and LR classifiers on the selected feature set for cardiovascular disease prediction reveals several remarkable findings (as shown in Tables 5 and 6. The performance metrics of SVM and Logistic Regression (LR) classifiers were evaluated in a comparative analysis of a stroke prediction dataset using a variety of validation methods (80-20 split, 5-fold CV, 10-fold CV) and sample sizes (original, 358, and 592). Consistent with the results, the two classifiers demonstrated comparable performance, with only minimal variations. Both classifiers maintained consistently high levels of accuracy, with SVM marginally surpassing LR in the 80-20 split for smaller sample sizes. Particularly in smaller sample sizes, LR exhibited a generally higher level of specificity, whereas SVM exhibited superior sensitivity, particularly in the 80-20 split for the smallest sample size of 358. However, SVM demonstrated marginal improvements in smaller samples, while both classifiers maintained identical precision on the original data. Minor differences were observed in the F1-score comparisons, with LR having a modest advantage in the original dataset's 80-20 split, but SVM performing better in smaller samples. These findings indicate that, although both classifiers are highly effective in stroke prediction, SVM may be more suitable for scenarios that highlight sensitivity, where LR is more appropriate when specificity is a critical factor. The graphical presentation is shown in Fig. 7.

**Table 5:** Performance of SVM Classifier on Selected Feature Set for Stroke Prediction Dataset

| Metrics | 80-20 split (Original data) | 5-CV (Original data) | 10-CV (Original data) | 80-20 split (sample size 358) | 5- fold CV (sample size 358) | 10- fold CV (sample size 358) | 80-20 split (sample size 592) | 5- fold CV (sample size 592) | 10fold CV (sample size 592) |
|---|---|---|---|---|---|---|---|---|---|
| accu-racy | 0.74 | 0.78 | 0.78 | 0.76 | 0.71 | 0.73 | 0.76 | 0.77 | 0.77 |
| specific-ity | 0.73 | 0.73 | 0.75 | 0.65 | 0.65 | 0.64 | 0.74 | 0.71 | 0.70 |
| sensitiv-ity | 0.81 | 0.78 | 0.78 | 0.88 | 0.71 | 0.73 | 0.81 | 0.77 | 0.77 |
| preci-sion | 0.94 | 0.78 | 0.78 | 0.78 | 0.73 | 0.73 | 0.81 | 0.77 | 0.77 |
| F1-score | 0.81 | 0.78 | 0.78 | 0.76 | 0.72 | 0.72 | 0.78 | 0.77 | 0.77 |

**Table 6:** Performance of LR Classifier on Selected Feature Set for Stroke Prediction Dataset

| Metric | 80-20 split (Original data) | 5-CV (Original data) | 10-CV (Original data) | 80-20 split (sample size 358) | 5- fold CV (sample size 358) | 10- fold CV (sample size 358) | 80-20 split (sample size 592) | 5- fold CV (sample size 592) | 10fold CV (sample size 592) |
|---|---|---|---|---|---|---|---|---|---|
| accuracy | 0.75 | 0.78 | 0.78 | 0.73 | 0.71 | 0.73 | 0.76 | 0.76 | 0.76 |
| specificity | 0.74 | 0.74 | 0.74 | 0.67 | 0.66 | 0.65 | 0.75 | 0.71 | 0.71 |
| sensitivity | 0.81 | 0.78 | 0.78 | 0.78 | 0.71 | 0.71 | 0.78 | 0.76 | 0.76 |
| precision | 0.94 | 0.78 | 0.78 | 0.74 | 0.72 | 0.71 | 0.81 | 0.76 | 0.76 |
| F1-score | 0.82 | 0.78 | 0.78 | 0.74 | 0.71 | 0.71 | 0.78 | 0.76 | 0.76 |

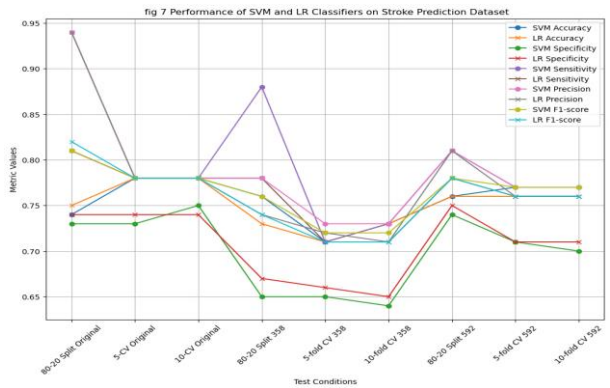


**Fig. 7:** Comparison of SVM and LR Performance Metrics on the Stroke Dataset.

To enhance clinical reliability and model evaluation rigour, we incorporated two additional metrics: the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The AUC-ROC quantifies the model's ability to distinguish between classes across thresholds. Our results show that the Support Vector Machine (SVM) classifier achieved an AUC-ROC of 0.965, indicating outstanding discrimination. The Logistic Regression (LR) model followed closely with an AUC of 0.937. The AUCROC curve for the Cleveland dataset is shown in Figure 8. Additionally, the AUC-ROC curve for the comprehensive dataset is shown in Figure 9, and the AUC-ROC curve for the stroke dataset is presented in Figure 10.

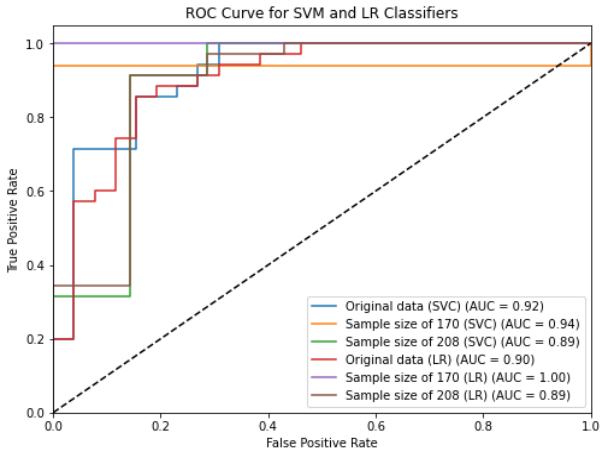


**Fig. 8:** ROC curves of SVM and LR on original and sampled datasets with AUC values of the Cleveland dataset.

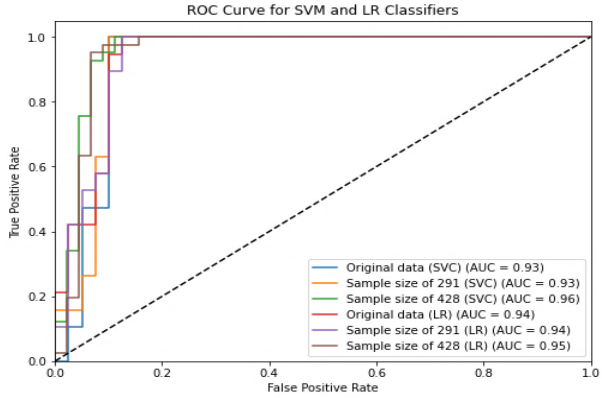


**Fig. 9:** ROC curves of SVM and LR on original and sampled datasets with AUC values of the Comprehensive dataset.

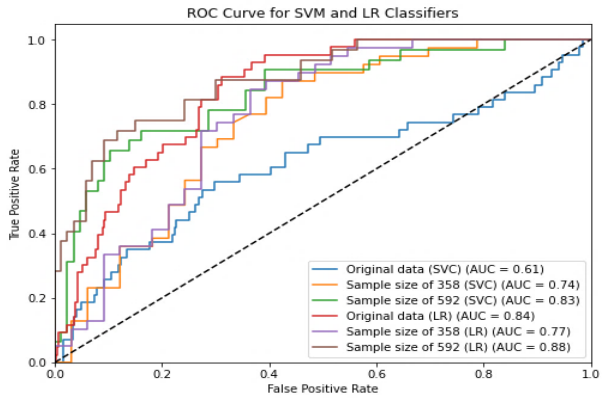


**Fig. 10:** ROC curves of SVM and LR on original and sampled datasets with AUC values of the Stroke dataset.

### 4.3. Shap-based analysis of feature importance

Figures 11 to 13 illustrate the SHAP-based interpretability of our models. We observed that in smaller samples, SHAP values for key features such as cholesterol, age, and blood pressure fluctuated significantly across test samples, suggesting lower interpretability reliability. In contrast, models trained on larger samples exhibited more stable SHAP feature attributions, reinforcing the idea that sample size impacts the consistency of interpretability.

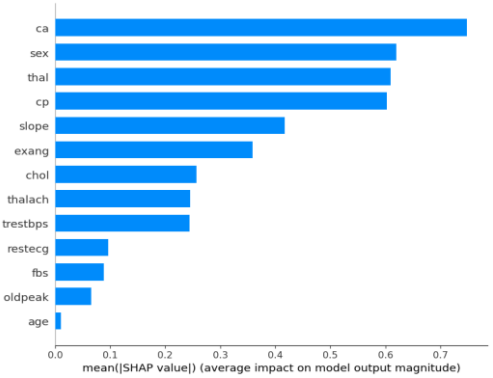


**Fig. 11:** SHAP bar plot showing the top features influencing the Cleveland dataset model predictions.

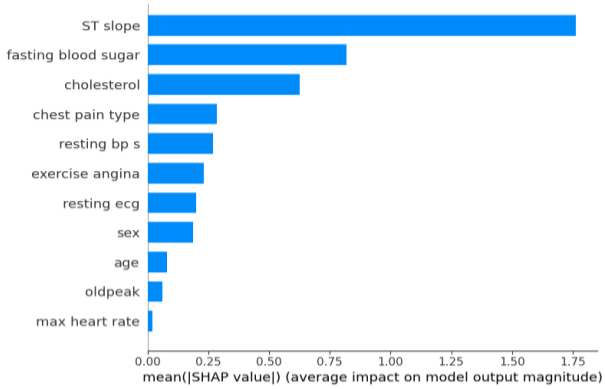


**Fig. 12:** SHAP bar plot showing the top features influencing the Comprehensive dataset model predictions.

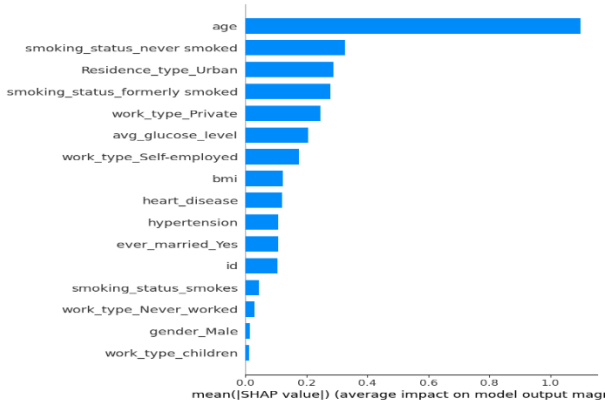


**Fig. 13:** SHAP bar plot showing the top features influencing the Stroke dataset model predictions.

## 5. Discussion

The findings of this study emphasise the crucial role of sampling strategy in influencing both the predictive accuracy and interpretability of machine learning models for clinical diagnosis. Support Vector Machine (SVM) consistently demonstrated higher accuracy compared to Logistic Regression (LR) across all three datasets. On the Cleveland dataset, SVM achieved an accuracy of 93% with 10-fold cross-validation, while LR achieved 84%. Similar patterns are seen in the Comprehensive dataset, where SVM attained 94% accuracy, while LR declined to 80%. In the Stroke dataset, both models performed similarly, but SVM marginally outperformed LR in sensitivity (e.g., 88% vs. 78%) on the smallest sample size of 358). Mutual Information-based feature selection identified dataset-specific key features such as 'thal', 'ca', and 'oldpeak' for Cleveland and 'age', 'work type', and 'BMI' for Stroke, highlighting the models' ability to identify clinically relevant predictors. SHAP analysis proved that larger sample sizes produce more consistent and interpretable feature importance explanations, while smaller samples show fluctuating SHAP values, particularly for variables like cholesterol and age. Additionally, AUC-ROC scores reinforced model reliability, with SVM reaching 0.965 compared to 0.937 for LR on the Cleveland dataset. Overall, these results demonstrate that adequate sampling improves not only prediction accuracy but also the stability and clinical trustworthiness of interpretability tools such as SHAP.

## 6. Conclusion

The study signifies the superior performance and reliability of the Support Vector Machine classifier over Logistic Regression in the context of both heart disease and stroke prediction tasks. SVM achieves higher classification metrics across different sample sizes and validation techniques, and demonstrates greater stability and generalisation under rigorous evaluation scenarios. Our findings support SVM as a robust predictive model across sample sizes, but with limited interpretability. Logistic regression suggests more clinical transparency despite a

slight drop in accuracy. Future studies may explore hybrid models combining SVM with SHAP or federated learning for privacy-conscious deployment.

## Dataset diversity

The Cleveland, Comprehensive, and Stroke datasets differ in size, structure, and feature composition, but they mostly represent confined geographical and demographic contexts. The sampling technique improves model performance across sample sizes, but dataset diversity limits it. The Cleveland dataset has just 303 occurrences, mainly from a homogenous clinical group, which could limit its applicability to multi-ethnic communities. Clinical predictions could be biased when models are trained on data without minority or socioeconomic representation, raising ethical problems. Critical factors like 'thal', 'job type', and 'BMI' may interact differently among demographic groupings, affecting model fairness. To establish reasonable, inclusive, and generalisable CVD and stroke prediction models, future research should use bigger, multi-centre datasets with different populations. Accountability and prevention of healthcare inequities require ethical precautions, including bias detection, subgroup evaluation, and model interpretability transparency (SHAP values).

## Ethical considerations

This study adheres to the highest ethical standards in academic research and publication. The data utilised in this work are sourced from publicly available repositories and do not involve any direct human or animal subjects, thereby exempting it from institutional ethical review requirements. All procedures performed in this research comply with relevant guidelines and regulations concerning data privacy and integrity. The authors affirm that this manuscript is original, has not been published previously, and is not under consideration for publication elsewhere. No conflicts of interest, whether personal, financial, or professional, exist among the authors that could influence the outcomes or interpretations of this research.

## References

[1] Wu R, Peters W and Morgan M W 2002. The next generation of clinical decision support: linking evidence to best practice. J. Healthcare Inf. Manag. 16(1): 50–55.
[2] Thuraisingham B.. 2000. A primer for understanding and applying data mining. IT Prof. 2(1): 28–31 https://doi.org/10.1109/6294.819936.
[3] Rajkumar A, and Sophia R G. 2010. Diagnosis of heart disease using a data mining algorithm. Global J. Comput. Sci. Technol. 10: 38–43.
[4] Anbarasi M, Anupriya E and Iyengar N C S N 2010. Enhanced prediction of heart disease with feature subset selection using a genetic algorithm. Int. J. Eng. Sci. Technol. 2: 5370–5376.
[5] Palaniappan S and Awang R, 2008. Intelligent heart disease prediction system using data mining techniques.Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. pp. 108–115 https://doi.org/10.1109/AICCSA.2008.4493524.
[6] Tripoliti E, Papadopoulos E, Karanasiou T G, Naka G S and Fotiadis K K D I. 2017 Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput. Struct. Biotechnol. J. 15:26 47. https://doi.org/10.1016/j.csbj.2016.11.001.
[7] Dash S R, Syed A S and Samantaray A. 2018. Filtration and classification of ECG signals. Handbook Res. Inf. Secur. Biomed. Signal Process. 72–94. https://doi.org/10.4018/978-1-5225-5152-2.ch005.
[8] Urbanowicz R J, Meeker M, La Cava W, Olson R S and Moore J H 2018 Relief-based feature selection: Introduction and review. J. Biomed. Inform. 85: 189–203. https://doi.org/10.1016/j.jbi.2018.07.014.
[9] Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Ser. B: Stat. Methodol. 58: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[10] Santhanam T and Ephzibah E P 2013 Heart disease classification using PCA and feed forward neural networks. Proc. 1st Int. Conf. MIKE pp. 90–99. https://doi.org/10.1007/978-3-319-03844-5_10.
[11] Ziasabounchi N and Askerzade I N 2014 A comparative study of heart disease prediction based on principal component analysis and clustering methods. Turkish J. Math. Comput. Sci. 16: 18.
[12] Akhil Jabbar M, Deekshatulu B L and Chandra P 2016 Prediction of heart disease using random forest and feature subset selection. Proc. 6th Int. Conf. IBICA pp. 187–196. https://doi.org/10.1007/978-3-319-28031-8_16.
[13] Kavitha R and Kannan E, 2016 An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. in Proc. ICETETS pp. 1–5. https://doi.org/10.1109/ICETETS.2016.7603000.
[14] Khateeb N and Usman M 2017. Efficient heart disease prediction system using K-nearest neighbour classification technique. Proc. Int. Conf. Big Data Internet Things pp. 21–26. https://doi.org/10.1145/3175684.3175703.
[15] Gokulnath C B and Shantharajah S P 2019 An optimised feature selection based on genetic approach and support vector machine for heart disease. Cluster Comput. 22: 14777–14787. https://doi.org/10.1007/s10586-018-2416-4.
[16] Sata M and Elkonca F 2020 A comparison of classification performances between the methods of logistic regression and CHAID analysis accordance with sample size. Int. J. Contemp. Educ. Res. 7(2): 15–26. https://doi.org/10.33200/ijcer.733720.
[17] Jindal H, Agrawal S, Khera R, Jain R and Nagrath P 2021. Heart disease prediction using machine learning algorithms. IOP Conf. Ser: Mater. Sci. Eng. 1022(1): 012072. https://doi.org/10.1088/1757-899X/1022/1/012072.
[18] Boukhatem C, Youssef H N and Bou A. 2022 Heart disease prediction using machine learning. Proc. ASET pp. 1–6. https://doi.org/10.1109/ASET53988.2022.9734880.
[19] Kavya S M, Deepasindhu M, Nowshika B and Shijitha R. 2023 Heart Disease Prediction Using Logistic Regression. J. Coastal Life Med. 11: 573–579.
[20] Chaudhuri A K, Das S and Ray A. 2024 An Improved Random Forest Model for Detecting Heart Disease. Data-Centric AI Solutions Emerg. Technol. Healthcare Ecosyst. pp. 143–164. https://doi.org/10.1201/9781003356189-10.
[21] Takahara, M., Katakami, N., Hayashino, Y. et al. Different impacts of metabolic profiles on future risk of cardiovascular disease between diabetes with and without established cardiovascular disease: the Japan diabetes complication and Its Prevention Prospective Study 7 (JDCP study 7). Acta Diabetol 59, 57–65 (2022). https://doi.org/10.1007/s00592-021-01773-z.
[22] Janosi A, Steinbrunn W, Pfisterer M and Detrano R. 1988. Heart Disease UCI Machine Learning Repository.
[23] Smith J 2023 Heart Disease Dataset. https://www.kaggle.com/datasets/ johnsmith88/heart-disease-dataset/data Accessed: 2024-06-18.
[24] Fedesoriano 2023 Stroke Prediction Dataset. https://www.kaggle.com/datasets/ fedesoriano/stroke-prediction-dataset Accessed: 2024-06-18.