Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Sample Size–Driven Approach to Heart Disease Risk Prediction

Vijayalakshmi Sarraju[1], jayapal[2], Supreeti.kamilya[3]
[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, Lalpur, Ranchi, India.*
[3] *Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India.*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Introduction:
Predicting cardiovascular disease survival outcomes is a challenging clinical data analytics subject with practical implications.

Objectives:
This paper analyses the association between sample size and model performance, providing insights relevant to generating reliable predictions across three diverse datasets.

Methods:
We use filter-based mutual information gain to identify significant characteristics. The Mutual Information gain methodology computes the dependency between each predictor variable and the target outcome, enabling the identification of the attributes that provide the most predictive value. Unlike wrapper techniques, mutual information gain is appropriate for clinical prediction applications since it is computationally competent and scalable to massive data sets. This novel approach of mutual information gain with sample-based statistical validation ensures robust and interpretable model performance across varied population sizes. Machine learning models, such as the support vector machine (SVM) and logistic regression (LR), are utilised to analyse sample sizes and assess the model's efficiency.

Results:
Across all datasets, larger samples consistently increased accuracy by up to 10%, improved sensitivity by 5–8%, and enhanced specificity, creating the positive impact of statistically representative sample sizes on model generalisation.

Keywords: Cardiovascular disease (CVD), Support Vector Machine (SVM), Logistic Regression (LR), Sample size.

INTRODUCTION

Health professionals diagnose cardiovascular disease (CVD) using complex clinical and pathological data, often using numerous diagnostic tests, specialist opinions, and prolonged medical assessments[1]. Because of more prolonged diagnosis 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 important risk factors and a lack of consistency in diagnostic techniques.

The World Health Organisation indicates that cardiovascular disease impacts one-third of individuals in underdeveloped nations. American Heart Association: One-third of people suffer from cardiovascular illnesses. Predictive models can enhance diagnosis by employing diverse data combinations and expert insights. Using multiple data configurations and expert knowledge, prediction models can improve diagnosis. This prediction process requires 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 effective strategy for classifying cardiovascular disease in clinical data, to predict heart disease stages[4-5].

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine learning algorithms can predict cardiovascular disease by studying complex patterns and risk factors in large datasets. These technologies can quickly identify high-risk patients for personalized treatment. Medical history, genetics, lifestyle, and biomarkers are used to create accurate prediction models. These algorithms assist clinicians in identifying constant treatments and appropriate patient care to reduce cardiovascular disease mortality. Research will demonstrate how SVM and LR may predict cardiovascular disease. Each technique has advantages and disadvantages depending on the investigation's facts and goals. Logistic regression assesses age, cholesterol, and blood pressure as CVD risk variables. Individual risk factors affect patient therapy and professional decision-making. SVM kernels capture non-linear variable interactions and efficiently handle 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 significantly correlated to 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 gives a statistical and machine learning literature overview on CVD. Section 3 discusses the dataset description and methodology. Section 4 presents an analysis of the results. Section 5 represents a discussion of the study. The article concludes in Section 6.

LITERATURE SURVEY

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

Table: Literature Review Summary with Research Gaps

Study	Approach	Findings	Research Gap
Kavitha & Kannan (2016)[10]	PCA to minimize data size	Enhanced accuracy, reduced computation	No statistical feature validation; interpretability not discussed
Khateeb & Usman (2017)[11]	Evaluated NB, KNN, DT, and bagging on Cleveland dataset	92% with KNN	There is no discussion of feature relevance or statistical validation
Gokulnat & Shantharajah[12]	Genetic algorithm + ML models (SVM, etc.) for feature selection	88.34% with SVM	Computationally heavy; lacks a statistical foundation
Mehmet Sata et al. (2020)[13]	Studied LR vs CHAID with different sample sizes	CHAID improved with >1000 samples	Did not integrate with filter-based feature selection techniques

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

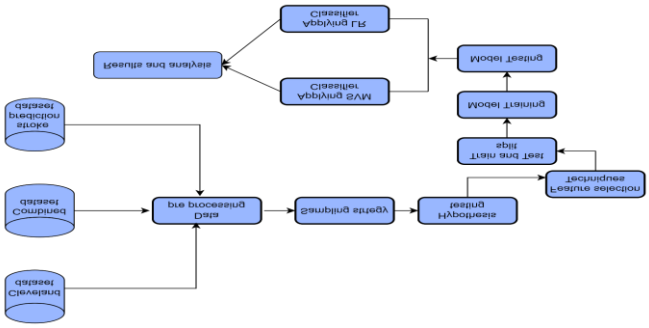
Research Article

Harshit Jindal et al. (2021)[14]	KNN, LR, NB with medical history	Effective classification	Lacks a deep evaluation of model behaviour under varying sample sizes
Chaimaa Boukhatem et al. (2022)[15]	MLP, SVM, RF, NB + feature selection and preprocessing	SVM: 91.67%	Does not analyse model stability across datasets or sample sizes
Kavya S M et al. (2023)[16]	Kaggle data; LR used for 10-year CHD risk	Multi-class classification	Limited comparison between sample sizes or filter-based methods
Arkadip Ray et al. (2024)[17]	ILR + ML classifier comparison	ILR: 83% accuracy	interpretability or visual diagnostics not assessed

OBJECTIVES

The objective is to analyse the impact of sample size on CVD prediction using filter-based variable selection techniques and machine learning models. Figure 1 depicts the workflow of the research framework for CVD prediction.

Fig. 1: Workflow diagram



METHODS

Dataset

This study utilizes three medical datasets to evaluate predictive modelling for cardiovascular and stroke risk: (i) the Cleveland Heart Disease dataset (UCI repository), (ii) a comprehensive heart disease dataset aggregated from Cleveland, Statlog, Hungary, Switzerland, and Long Beach VA sources, and (iii) a stroke prediction dataset from Kaggle. The Cleveland dataset contains 303 records and 13 clinical features such as 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 merges similar features across five data sources, resulting in a feature-rich, heterogeneous dataset with 1190 instances and 12 attributes. The stroke prediction dataset comprises 5110 entries and includes 11 features combining 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.

Data Preprocessing

Data preprocessing involved duplicate removal and outlier detection using the Z-score method, where data points with scores exceeding ±3 were excluded. The comprehensive dataset underwent **min-max normalisation** to rescale values to the [0, 1] range due to the presence of negative values, which could bias distance-based models. Normalisation was not applied to the Cleveland and stroke datasets, as their features were already on compatible scales. These preprocessing steps ensured consistent data quality across all datasets, enhancing model reliability. After cleaning, the final usable records were 302 (Cleveland), 918 (comprehensive), and 4228 (stroke).

3.3 Sampling Approach

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

To evaluate model robustness and generalisation, statistically valid sample sizes were drawn based on 95–99% confidence intervals, 50% population proportion, and a 5% margin of error. A t-test is employed to assess whether sample means differ significantly from population means. Results confirmed no significant differences, supporting that the sampled subsets were representative. The selected sample sizes included 170 and 280 for the Cleveland dataset, 291 and 428 for the comprehensive dataset, and 358 and 592 for the stroke dataset.

A filter-based feature selection technique utilising Mutual Information Gain (MIG) was employed to ascertain the most informative predictors. MIG is statistically found by quantifying the relationship between input characteristics and the target variable, enabling dependable selection. Its computational efficiency is especially appropriate for clinical datasets, where both interpretability and speed are essential. Two supervised learning models, Support Vector Machine (SVM) and Logistic Regression (LR), were developed utilising an initial 80-20 training-testing partition. To mitigate overfitting and enhance model generalisation, 5-fold and 10-fold cross-validation approaches were employed. These validation procedures guarantee that every data point is utilised in training and testing, yielding more dependable performance assessments. The assessment used essential measures, including accuracy, precision, sensitivity, specificity, and F1-score.

RESULTS

DATASET	MODEL	MAX ACCURACY	MAX F1-SCORE	BEST VALIDATION METHOD
CLEVELAND	SVM	0.93	0.93	10-FOLD CV
CLEVELAND	LR	0.94	0.93	80-20 SPLIT (N=170)
COMBINED	SVM	0.94	0.95	5-FOLD CV (N=428)
COMBINED	LR	0.93	0.92	80-20 SPLIT (N=291)
STROKE	SVM	0.78	0.78	5/10-FOLD CV
STROKE	LR	0.78	0.78	5/10-FOLD CV

AUC-ROC CURVE ANALYSIS

A classification model's ROC (Receiver Operating Characteristic) domain efficacy is measured by its Area Under the Curve (AUC). The ROC curve quantifies an essential element of the area under the curve that depicts the relationship between sensitivity and (1-specificity) at various classification thresholds. The numerical AUC value ranges from 0 to 1. Better model performance means a higher value. Values below 0.5 indicate poor model performance, whereas 0.5 indicates random performance. The AUC measures the classifier's likelihood of prioritising a randomly picked positive observation above a negative one. This technique provides relevant performance evaluations even for datasets with highly unbalanced classes.

In this study, we analysed the AUC-ROC curves for various datasets. The AUCROC curve for the Cleveland dataset is shown in Figure 2. Additionally, the AUC-ROC curve for the comprehensive dataset can be seen in Figure 3, and the AUC-ROC curve for the stroke dataset is presented in Figure 4.

According to curve analysis on the Cleveland, comprehensive, and stroke prediction datasets, Logistic Regression and Support Vector Machine algorithms are effective cardiovascular risk assessors. The Logistic Regression model consistently generates high AUC values of 0.84 to 1.00 across all datasets, demonstrating it can recognise cardiovascular event risk variables. The SVM model operates well, with AUC values from 0.61 to 0.96, although datasets, especially the stroke prediction dataset, oscillate. As ROC curves indicate, both models enable doctors to categorise risk and identify high-risk patients. Both models help assess and control cardiovascular risk, enhancing clinical decision-making and patient care. Logistic Regression is more consistent and performs better across datasets.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

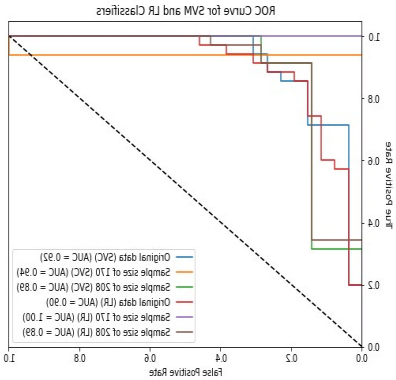


FIG:2

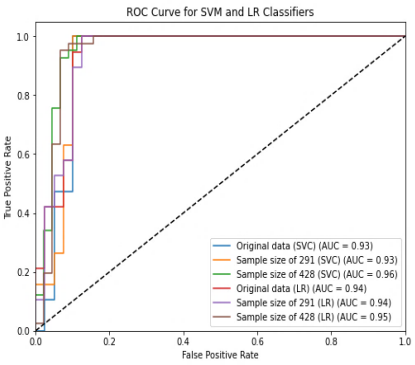


FIG: 3

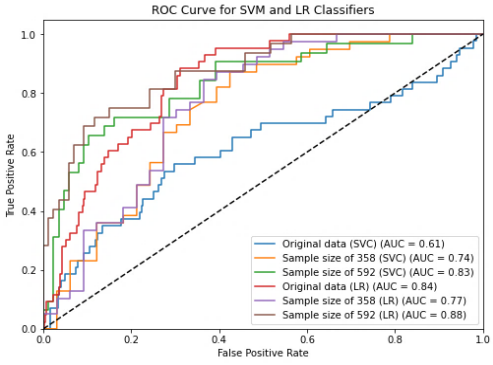


FIG:4

DISCUSSION

The comparative performance results indicate that the SVM classifier outperforms Logistic Regression (LR) in almost all metrics and dataset configurations. On the Cleveland dataset, SVM achieved an accuracy of up to 0.93 and continued high F1-scores (≥0.90) across all validation patterns, including 10-fold CV. Comparatively, LR's accuracy declined to 0.80 in sensitivity and precision. This performance gap is more prominent when smaller sample sizes (170 and 208) are used. On the combined dataset, SVM stabilised strong classification results (accuracy ≥0.92 and specificity up to 1.00), while LR's performance again declined under k-fold validation. A similar trend is evident in the Stroke dataset: although both models experience reduced accuracy (around 0.76–0.78), SVM consistently produces a better balance between sensitivity and specificity than LR, which exhibits more fluctuation, particularly in cross-validation. These observations suggest that SVM is less sensitive to sample size reduction and handles non-linear feature interactions better, leading to more reliable predictions across varied scenarios.

CONCLUSION

The study emphasises 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 varying dataset sizes and validation techniques and demonstrates greater stability and generalisation under rigorous evaluation scenarios. Simultaneously, logistic regression is a computationally efficient and interpretable baseline, but its efficacy reduces throughout cross-validation, especially with small datasets and increased data complexity. Consequently, for medical diagnostics, where accuracy, generalizability, and reliability are paramount, SVM is a more robust and preferable model. Future work could explore ensemble methods or hybrid architectures to enhance predictive reliability, especially in limited and imbalanced clinical data cases.

REFRENCES

[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
[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
[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
[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

Research Article

[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.
[9] Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Ser. B: Stat. Methodol. 58: 267–288.
[10] 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
[11] 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
[12] 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
[13] 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
[14] 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
[15] Boukhatem C, Youssef H N and Bou A. 2022 Heart disease prediction using machine learning. Proc. ASET pp. 1–6
[16] Kavya S M, Deepasindhu M, Nowshika B and Shijitha R. 2023 Heart Disease Prediction Using Logistic Regression. J. Coastal Life Med. 11: 573–579
[17] 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