Sri Lankan Journal of Applied Statistics, Vol (27-1)

Cost Sensitive and Explainable Machine Learning for Heart Condition Prediction on Imbalanced Data

K. A. M. N. Aryawansha[1,*], C. H. Magalla[1], and N. Hettiarachchi[2]

[1]Department of statistics, Faculty of science, University of Colombo, Colombo 03, Sri Lanka

[2]Tokyo Metropolitan Institute of Medical Science, 2-1-6 Kamikitazawa, Setagaya ku,Tokyo, Japan

*Corresponding author: maleeshaaryawansha@gmail.com

Received: 17th November, 2025/ Revised: 18th February, 2026/ Accepted: 02nd March, 2026/ Published: 31st March, 2026

ABSTRACT

Heart disease remains a leading cause of mortality worldwide, highlighting the need for predictive models that can reliably detect high risk individuals under severe class imbalance. This study presents a cost sensitive and explainable learning framework for heart condition prediction using machine learning, deep learning, and hybrid models, with particular emphasis on how dataset size influences both model effectiveness and model selection. The 2022 CDC Behavioral Risk Factor Surveillance System (BRFSS) dataset, consisting of 246,013 records with 40 features and 8.8% positive cases, was used to generate subsets ranging from 5K to the full dataset while preserving the original class distribution. Models were evaluated under Synthetic Minority Oversampling Technique (SMOTE) and Cost Sensitive Learning (CSL), and performance was assessed using recall, F1 score, and AUC–ROC. Across all model categories, CSL consistently outperformed SMOTE by improving minority class detection while maintaining data integrity. On the full dataset, logistic regression improved from a minority class F1 score of 0.320 with recall of 0.624 under SMOTE to an F1 score of 0.364 and recall of 0.782 under CSL, corresponding to an improvement of approximately 14% in F1 score. Random forest showed a stronger absolute gain, with F1 increasing from 0.173 to 0.359 and recall from 0.111 to 0.778 under CSL. Dataset size

effects were evident when analyzed using CSL results: for the 5K subset, logistic regression achieved a minority class F1 score of 0.389 and the multi layer perceptron achieved 0.407, whereas on the full dataset the best performing models were random forest and TabNet, with minority-class F1 scores in the range of 0.34–0.39, while the multi layer perceptron showed reduced generalization. These findings demonstrate that model performance is not constant across dataset sizes and that optimal prediction requires selecting the most suitable model within each paradigm. Finally, a quantum support vector machine is explored as a proof of concept to compare quantum and classical kernel behavior, highlighting future potential for quantum enhanced medical analytics.

Keywords: Heart condition prediction, Machine learning, Deep learning, Hybrid models, Quantum machine learning, Cost-sensitive learning

1 Introduction

Heart conditions, including cardiovascular diseases (CVD) and other heart related disorders, remain a leading cause of global mortality. CVD, which affects the heart or blood vessels, arises from a complex interplay of socioeconomic, metabolic, behavioral, and environmental risk factors. Approximately 17.9 million deaths annually are attributed to heart disease, reflecting its substantial public health burden (World Health Organization, 2025). Over the past three decades, the absolute number of CVD related deaths has increased; however, when adjusted for population age structure, mortality rates have declined. This reduction has been more pronounced in high income countries (HICs) compared to low- and middle income countries (LMICs), where over 80% of global CVD deaths occur (World Heart Federation, 2023b).

Heart conditions remain one of the leading causes of mortality worldwide, accounting for millions of deaths annually and imposing a disproportionate burden on low and middle income countries (World Heart Federation, 2023b). Early detection is essential, as cardiovascular abnormalities often remain undiagnosed until severe events such as heart attacks or strokes occur (World Health Organization, 2025; World Heart Federation, 2023a). Consequently, accurate and scalable predictive models are critical for improving early diagnosis and supporting preventive healthcare strategies.

Recent advances in machine learning and deep learning have enabled predictive models capable of capturing complex relationships among demographic, behavioral, and clinical risk factors associated with heart conditions (Garavand et al., 2022; Ahmad and Polat, 2023). However, medical datasets are typically characterized by high dimensionality and severe class imbalance, where positive heart condition cases are underrepresented. This imbalance often leads to inflated overall accuracy while minority class recall and F1 performance remain poor, limiting clinical applicability (Natarajan et al., 2024). To address class imbalance, data level techniques based on synthetic sample generation, such as SMOTE based approaches, have been widely adopted.

While these techniques may improve minority class representation, they can distort the original data distribution and introduce noise, potentially reducing medical validity in healthcare applications (Fernández et al., 2018). Algorithm level strategies, particularly cost sensitive learning, offer an alternative by assigning higher importance to clinically critical cases without modifying the underlying data distribution, thereby preserving the integrity of real-world medical data (Elkan, 2001). Despite this, systematic evaluations comparing synthetic oversampling and cost sensitive learning across varying dataset sizes remain limited.

Another important but underexplored factor in heart condition prediction is dataset size. In real world healthcare settings, data availability varies substantially due to geographic, demographic, and institutional constraints, especially in resource limited regions (World Heart Federation, 2023b). Nevertheless, most existing studies evaluate models using a single dataset size, providing limited insight into how model performance, generalization capability, and model suitability vary as data scale changes (Garavand et al., 2022). In addition, the black box nature of many machine learning and deep learning models restricts their adoption in clinical practice, highlighting the need for explainable AI techniques that provide transparent and medically interpretable predictions (Tjoa and Guan, 2021).

The core problem addressed in this study is the lack of a unified evaluation framework that jointly accounts for dataset size variability, class imbalance handling, and model interpretability in heart condition prediction. Without such a framework, reported model performance cannot be reliably generalized across real world healthcare scenarios where data characteristics vary substantially.

To address this problem, this study systematically evaluates machine learning, deep learning, and hybrid models across multiple dataset sizes using cost sensitive learning to handle class imbalance while preserving original data characteristics. Explainable AI techniques are incorporated to enhance transparency and clinical interpretability, and quantum machine learning is explored as a feasibility oriented extension to examine emerging trends in healthcare analytics, consistent with recent investigations into quantum approaches for medical data analysis (Babu et al., 2024).

2 literature review

Understanding and predicting heart conditions has been a longstanding objective in cardiovascular research. Early epidemiological studies, most notably the Framingham Heart Study (FHS), established the concept of cardiovascular risk factors and identified key contributors such as smoking, hypertension, and cholesterol levels, forming the foundation of modern cardiac risk assessment models (Mahmood et al., 2014; Hajar, 2017). Subsequent research expanded this understanding by identifying emerging lifestyle and environment-related risk factors, including socioeconomic status, sleep irregularities, and

behavioral patterns, reflecting the evolving and multifactorial nature of heart conditions (Adhikary et al., 2022; Javaheri et al., 2019; Sumida et al., 2019).
Building on risk factor identification, traditional predictive models such as logistic regression and the Framingham Risk Score (FRS) were developed to estimate cardiovascular risk using a limited set of clinical variables (Truett et al., 1967; Wilson et al., 1998). While these models offered interpretability and clinical usability, their reliance on linear assumptions and restricted feature sets limited their ability to capture complex interactions among risk factors (D'Agostino et al., 2008). These limitations motivated the adoption of machine learning (ML) and deep learning (DL) techniques, which leverage data driven learning to model nonlinear relationships in high dimensional health data (Chicco and Jurman, 2020).
Recent studies applying ML and DL to heart disease prediction have reported high predictive accuracy using algorithms such as Random Forest, XGBoost, Multi Layer Perceptrons (MLP), and TabNet (Bhatt et al., 2023; Pillai, 2022). However, many of these studies rely on small or artificially balanced datasets, which limits their generalizability to real world clinical settings where class imbalance is intrinsic. Moreover, performance is often evaluated primarily using accuracy, despite its inadequacy in imbalanced medical datasets where minority class detection is clinically critical.
Data quality and preprocessing play a central role in the reliability of predictive models. Medical datasets frequently contain heterogeneous variable types, high dimensionality, redundancy, and noise, which necessitate careful encoding, normalization, and feature handling (Benhar et al., 2020; Natarajan et al., 2024). Although techniques such as feature selection and dimensionality reduction can improve model efficiency, excessive preprocessing may suppress clinically meaningful variability. Similarly, outlier detection methods such as isolation forests are commonly used to identify anomalous observations, yet indiscriminate removal of outliers in medical data risks eliminating rare but clinically significant cases.
Class imbalance remains one of the most persistent challenges in heart condition prediction. Data level approaches based on synthetic sample generation, including SMOTE and its variants, have been widely adopted and shown to improve recall in minority classes (Tompra et al., 2024; Albert et al., 2023). However, synthetic oversampling alters the original data distribution and may compromise the physiological validity of learned patterns, particularly in population-level health surveys. Algorithm level strategies, especially cost sensitive learning (CSL), provide an alternative by incorporating class importance directly into the learning process without modifying the data. CSL has demonstrated improved minority class detection while preserving data integrity, making it particularly suitable for medical applications where authenticity of observations is critical.
Another underexplored but practically significant factor in heart condition prediction is dataset size. Existing studies typically evaluate models on a single dataset size, despite real world healthcare data varying substantially across

institutions and regions. Small datasets may favor simpler or more flexible models, whereas larger datasets enable better generalization but may expose limitations in certain algorithms. The lack of systematic evaluation across multiple sample sizes restricts understanding of how model suitability changes with data scale, highlighting an important gap in current literature.

In addition to predictive performance, interpretability is essential for clinical adoption of ML and DL models. The black box nature of many advanced algorithms limits clinician trust and hinders decision support. Explainable AI (XAI) techniques, particularly SHAP, have been shown to improve transparency by quantifying feature contributions and providing clinically interpretable explanations (Loh et al., 2022; Athanasiou et al., 2020). Integrating XAI into predictive pipelines enhances accountability and supports informed medical decision making.

More recently, Quantum Machine Learning (QML) has emerged as a promising extension of classical ML. Studies exploring quantum enhanced algorithms such as Quantum Support Vector Machines have reported modest improvements in accuracy and computational efficiency in heart disease prediction tasks (Babu et al., 2024). While practical deployment remains constrained by hardware limitations, QML represents a forward looking direction for scalable and efficient medical analytics.

Overall, existing literature demonstrates substantial progress in heart condition prediction using ML and DL; however, key gaps remain. These include limited evaluation across varying dataset sizes, overreliance on synthetic data generation for imbalance handling, insufficient emphasis on clinically relevant metrics, and restricted integration of interpretability. Addressing these gaps requires a unified framework that evaluates model performance across data scales, prioritizes algorithm level imbalance handling, and ensures transparency and clinical validity objectives that form the basis of the present study.

3 Methodology

This study adopts a structured analytical framework to investigate heart condition prediction under data imbalance and varying sample sizes. The methodology is designed to evaluate how different modeling paradigms machine learning (ML), deep learning (DL), hybrid approaches, and quantum machine learning (QML), behave when dataset scale and class imbalance are explicitly considered.

3.1 Data Source and Preprocessing

The dataset was obtained from Kaggle and compiled from the 2022 Behavioral Risk Factor Surveillance System (BRFSS) survey conducted by the Centers for Disease Control and Prevention. It consists of 246,013 records and 40 features, including both categorical and numerical variables, with no missing

values. Duplicate records were identified and removed. To ensure a clinically consistent target definition, heart condition related variables were merged to create a binary outcome variable indicating the presence or absence of a heart condition.
Categorical variables were encoded using binary, one hot, or ordinal encoding based on their semantic structure, while numerical features were standardized. Outlier detection was performed using Isolation Forest, identifying approximately 5% potential anomalies; however, no outliers were removed, as excluding clinically plausible cases without domain validation could compromise medical relevance.

3.2 Sample Size Regeneration

To examine the effect of dataset size on model behavior, multiple subsets were recreated from the original dataset using reproducible random sampling. Subsets of 5,000, 10,000, 20,000, 50,000, 100,000, and the full dataset were generated while preserving the original class imbalance ratio (approximately 10.4). Extremely small datasets were avoided, as they are unlikely to support reliable learning in ML and DL models.

3.3 Handling Class Imbalance

Class imbalance is a central challenge in heart condition prediction. Two imbalance handling strategies were deliberately selected: Synthetic Minority Oversampling Technique (SMOTE) and Cost Sensitive Learning (CSL). SMOTE was included as a representative data level approach due to its widespread use in medical prediction studies. While several SMOTE variants exist (e.g., ADASYN, SMOTE-ENN), the standard formulation was chosen to avoid introducing multiple synthetic mechanisms that could confound interpretability and clinical validity.
CSL was adopted as the primary strategy because it operates at the algorithmic level, preserving the original data distribution while penalizing clinically critical misclassifications. Four weighting schemes were evaluated: imbalance ratio, inverse imbalance ratio, inverse class frequency, and logarithmic inverse class frequency. Other imbalance-handling approaches, such as undersampling or ensemble resampling, were not used because they either discard valuable patient data or introduce additional model complexity beyond the study's scope.

3.4 Machine Learning, Deep Learning, and Hybrid Models

Five classical ML models were employed: Logistic Regression, Random Forest, XGBoost, LightGBM, and Naïve Bayes. These models represent linear, probabilistic, and ensemble based learning paradigms and were selected to

capture diverse learning behaviors under imbalance and scale variation. Two DL models were used: a Multilayer Perceptron (MLP) and TabNet. MLP was included to evaluate fully connected neural learning, while TabNet was selected for its suitability to tabular medical data and built-in interpretability via sequential attention.

Hybrid models were introduced to assess whether combining complementary learning mechanisms improves robustness. A Random Forest–TabNet hybrid was constructed by using Random Forest for feature importance-driven selection followed by TabNet based learning. Additionally, a stacking ensemble was implemented using heterogeneous base learners with Logistic Regression as the meta learner to enhance generalization.

Hyperparameter optimization for selected best performing models was conducted using Optuna to ensure fair and consistent model tuning across dataset sizes.

3.5 Evaluation Strategy

Given the strong class imbalance, overall accuracy was not treated as the primary performance metric. Model evaluation emphasized recall, precision, F1-score for the minority class, and AUC–ROC, reflecting the clinical importance of detecting true heart condition cases while controlling false alarms.

3.6 Explainable AI

To address the black box limitation of ML and DL models, Shapley Additive Explanations (SHAP) were applied to interpret model predictions. SHAP was used to identify globally important features and to explain individual predictions, enhancing transparency and clinical trustworthiness.

3.7 Quantum Machine Learning: Proof of Concept

Quantum Machine Learning was included as an exploratory component rather than a competitive benchmark. A Quantum Support Vector Machine (QSVM) with a predefined quantum kernel was implemented using Qiskit to examine how cardiac related data behave in a quantum feature space. No custom quantum circuits were designed, and model complexity was intentionally limited due to current hardware constraints. The objective was to provide a proof of concept demonstration of feasibility rather than to claim performance superiority over classical models.

4 Results and Discussion

4.1 Exploratory Data Analysis Results

Exploratory Data Analysis (EDA) was conducted on the full dataset (n =

246,013) to identify statistically and practically relevant patterns that would inform preprocessing, imbalance handling, and model selection decisions.

Association and Feature Relevance

Chi-square tests indicated that all categorical variables were significantly associated with the response variable ($p < 0.001$). However, statistical significance alone was not used for feature prioritization due to the large sample size. Therefore, Cramér's V was employed to quantify the strength of association. The strongest categorical predictors were *GeneralHealth* (0.243), *AgeCategory* (0.242), *ChestScan* (0.215), *DifficultyWalking* (0.199), and *RemovedTeeth* (0.191). These findings justified retaining all features during model development while giving special attention to high impact variables during interpretability analysis using SHAP. The presence of multiple moderately associated variables further supported the use of ensemble and nonlinear models capable of capturing interaction effects beyond linear relationships.

Multicollinearity Assessment

Correlation analysis among numerical variables revealed strong correlation between *WeightInKilograms* and *BMI*, while other numerical features showed no high multicollinearity. As tree based models are robust to multicollinearity, both variables were retained. However, this observation supported the inclusion of regularized models (Logistic Regression with class weighting) to mitigate potential instability in linear estimation.

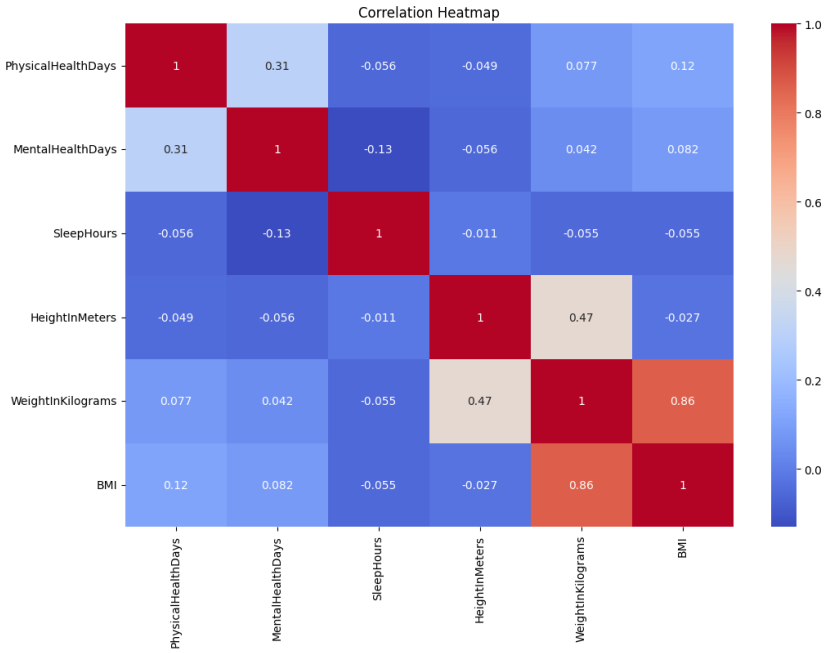

Fig. 1: Correlation between numerical variables

Class Imbalance Characteristics

The response variable exhibited severe imbalance (Class 0: 91.2%, Class 1: 8.8%), corresponding to an imbalance ratio of 10.4. This imbalance directly motivated the comparative evaluation of algorithm level (Cost Sensitive Learning) and data level (SMOTE) strategies. The imbalance severity also justified prioritizing recall and F1-score over overall accuracy during model evaluation.

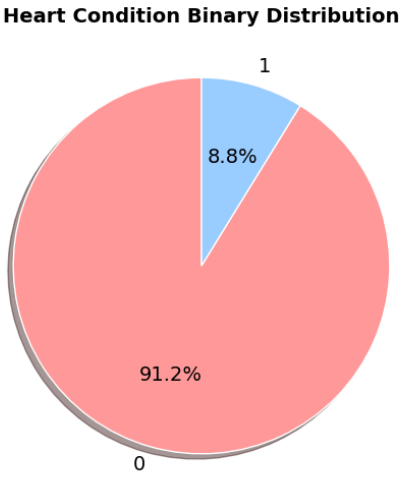

Fig. 2: Binary distribution of heart condition classes

Nonlinear Risk Patterns

Age stratified prevalence analysis revealed an accelerating increase in heart condition risk beyond mid life, indicating nonlinear risk growth.

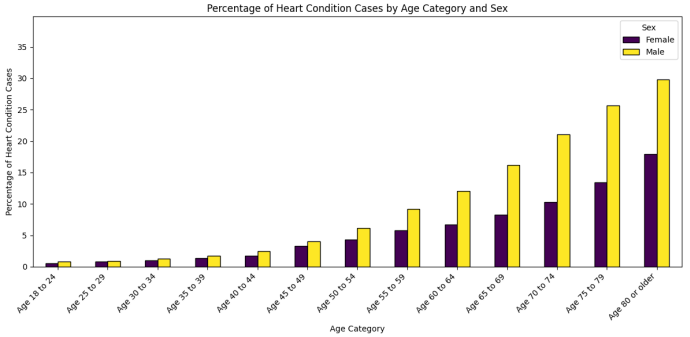

Fig. 3: Heart Disease Rates Across Different Age Demographics and Gender

Similarly, BMI related prevalence increased disproportionately in higher categories. These nonlinear patterns provided empirical justification for deploy-

ing ensemble (Random Forest, XGBoost, LGBM) and deep learning (TabNet, MLP) models, which are capable of modeling complex interactions that traditional linear models may fail to capture.

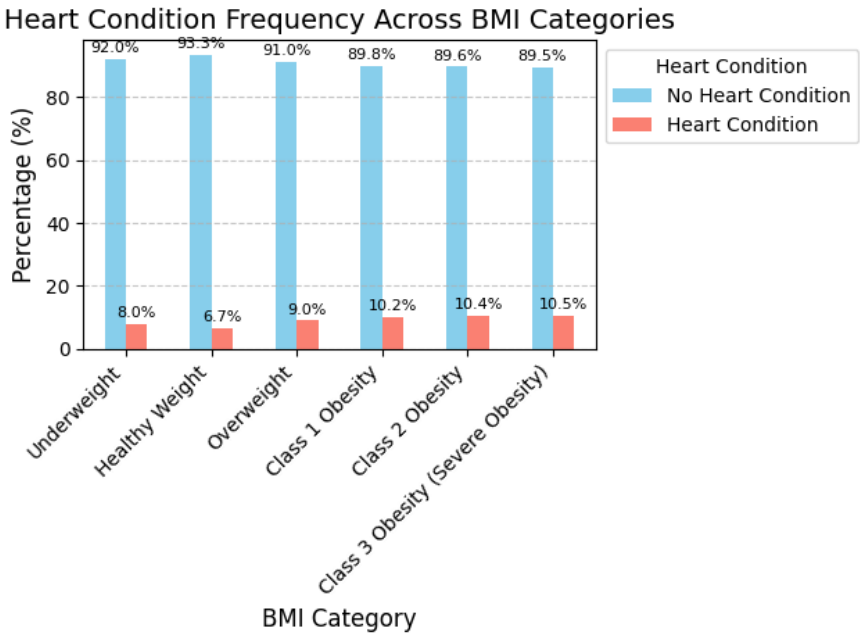

Fig. 4: Heart condition frequency across BMI categories, showing increased prevalence with higher BMI levels.

Lifestyle and Clinical Indicators

Smoking status demonstrated stronger association with heart conditions (Cramér's V = 0.106) compared to alcohol consumption (Cramér's V = 0.084) and e-cigarette usage (Cramér's V = 0.032). The moderate yet consistent strength of lifestyle and functional limitation variables reinforced the need for interpretable modeling, motivating the integration of SHAP based explanation methods to assess feature contributions at both global and local levels.

Table 1: Cramér's V Values for Lifestyle Related Categorical Variables

Variable	Cramér's V
Smoker Status	0.106
Alcohol Drinkers	0.084
E-Cigarette Usage	0.032

Implications for Modeling Strategy

Overall, EDA findings informed three major modeling decisions:

- Adoption of cost sensitive learning due to severe class imbalance.
- Use of nonlinear and ensemble architectures to capture accelerating and interaction based risk structures.
- Integration of SHAP to interpret clinically influential variables identified during association analysis.

Thus, EDA served not only as a descriptive step but as a structured decision making foundation guiding preprocessing, imbalance handling, model selection, and interpretability design.

4.2 Advance Modeling Results

This section reports the performance of machine learning (ML), deep learning (DL), and hybrid models under three training strategies: (i) standard training on the imbalanced dataset, (ii) SMOTE based oversampling, and (iii) cost sensitive learning (CSL). Models were evaluated using test accuracy, F1-score, recall, and AUC–ROC, with emphasis on minority class (Level 1: heart condition) performance because accuracy can be inflated under severe imbalance. Although six dataset sizes were evaluated (5K, 10K, 20K, 50K, 100K, and the full dataset), results are summarized for 5K and the full dataset due to page limits; full results for all sizes are provided in the appendix.

Why imbalance handling focused on SMOTE and CSL?

Class imbalance in medical screening is commonly handled using data level approaches (random over/under sampling, SMOTE variants such as ADASYN/SMOTE-ENN, and hybrid cleaning methods) or algorithm level approaches (class weights, focal loss, threshold moving, and cost sensitive training). In this study, SMOTE and CSL were selected to represent these two families because they are (i) widely used baselines in medical classification, (ii) can be implemented consistently across ML, DL, and hybrid models, and (iii) conceptually contrasting: SMOTE changes the training distribution using synthetic samples, while CSL preserves the original data and changes the misclassification penalties. This allowed a controlled comparison between "data level" and "algorithm level" imbalance handling within the paper constraints.

Model fitting on the original dataset

When trained on the original imbalanced dataset, most models achieved high accuracy (around 0.91) but performed poorly in detecting Level 1 cases. For example, Random Forest achieved test accuracy above 0.91 but minority recall as low as 0.03 (full dataset), and several models showed minority recall below 0.15. This confirms that standard training caused models to favor the majority class (Level 0), inflating accuracy while missing clinically critical

heart condition cases. Naïve Bayes was an exception: it achieved lower overall accuracy (0.76) but higher minority recall (0.64) and improved minority F1 (0.33), indicating increased sensitivity at the cost of specificity

Table 2: Model performance comparison on 5K and full, imbalanced datasets

Metric	Logistic Regression	Random Forest	XGBoost	LightGBM	Naive Bayes	TabNet	MLP	RF+TabNet	Stacking Model
5K Sample Size									
Test Accuracy	0.909	0.912	0.904	0.900	0.604	0.888	0.900	0.900	0.920
AUC-ROC	0.8594	0.8273	0.8021	0.8183	0.7138	0.7398	0.7703	0.7703	0.500
F1-Score (Level 0)	0.9520	0.9540	0.9491	0.9471	0.7321	0.9403	0.9500	0.9500	0.9600
F1-Score (Level 1)[1]	0.1495	0.0000	0.1429	0.0741	0.2414	0.0870	0.0000	0.0000	0.0000
Recall (Level 0)	0.9847	0.9967	0.9792	0.9792	0.5913	0.9837	1.000	1.000	1.000
Recall (Level 1)[1]	0.0941	0.0000	0.0941	0.0471	0.7412	0.0512	0.0000	0.0000	0.0000
Original Dataset									
Test Accuracy	0.9124	0.9110	0.9108	0.9122	0.7667	0.9137	0.910	0.9146	0.910
AUC-ROC	0.8437	0.8294	0.8410	0.8464	0.7851	0.8482	0.830	0.8486	0.500
F1-Score (Level 0)	0.9538	0.9533	0.9529	0.9538	0.8587	0.9544	0.950	0.9550	0.950
F1-Score (Level 1)[1]	0.1778	0.0630	0.1636	0.1409	0.3315	0.2031	0.180	0.1687	0.0000
Recall (Level 0)	0.9920	0.9976	0.9910	0.9943	0.7787	0.9896	0.990	0.9931	1.000
Recall (Level 1)[1]	0.1056	0.0333	0.0971	0.0801	0.6443	0.1253	0.110	0.0987	0.0000

[a]Metrics correspond to minority class (Level 1) performance.

Model fitting with SMOTE

Applying SMOTE increased minority class detection in several models but generally reduced overall accuracy. For example, Logistic Regression improved minority recall from about 0.10 (imbalanced training) to 0.62 (full dataset), and TabNet improved minority F1 from 0.20 to 0.26. DL models (MLP/TabNet) benefited more noticeably on small datasets, reaching minority recall around 0.40 under 5K. However, test accuracy commonly dropped to roughly 0.76–0.84, and performance improvements were not consistent for ensemble models. Since SMOTE introduces synthetic minority samples, it can distort the original data distribution an important concern in medical data where physiological authenticity matters. Therefore, SMOTE was treated as a comparison baseline rather than the primary approach.

Table 3: Model performance comparison for 5K sample size and original dataset with SMOTE

Metric	Logistic Regression	Random Forest	XGBoost	LightGBM	Naive Bayes	TabNet	MLP	RF+TabNet	Stacking Model
5K Sample Size									
Test Accuracy	0.7710	0.9000	0.8950	0.9030	0.4210	0.8387	0.82	0.7907	0.91
AUC-ROC	0.6896	0.7479	0.7540	0.7663	0.5980	0.7397	0.7742	0.7102	0.5484
F1-Score (Level 0)	0.8652	0.9471	0.9440	0.9485	0.5515	0.9079	0.95	0.8765	0.95
F1-Score (Level 1)[1]	0.2392	0.0741	0.1600	0.1565	0.1834	0.3529	0.31	0.3144	0.18
Recall (Level 0)	0.8068	0.9835	0.9715	0.9813	0.3908	0.8856	0.87	0.8276	0.98
Recall (Level 1)[1]	0.4045	0.0449	0.1124	0.1011	0.7303	0.4286	0.40	0.4675	0.12
Original Dataset									
Test Accuracy	0.7620	0.9043	0.9089	0.9104	0.5930	0.8411	0.09	0.8289	0.9
AUC-ROC	0.7777	0.8172	0.8324	0.8366	0.7145	0.7873	0.4064	0.7835	0.5735
F1-Score (Level 0)	0.8557	0.9492	0.9518	0.9526	0.7222	0.9109	0.00	0.9027	0.95
F1-Score (Level 1)[1]	0.3201	0.1725	0.1660	0.1791	0.2391	0.2644	0.16	0.2938	0.24
Recall (Level 0)	0.7756	0.9825	0.9886	0.9894	0.5812	0.8908	0.00	0.8697	0.98
Recall (Level 1)[1]	0.6242	0.1112	0.1010	0.1089	0.7122	0.3252	1.00	0.4051	0.17

[a]Metrics correspond to minority class (Level 1) performance.

Model fitting with Cost Sensitive Learning:

Cost Sensitive Learning (CSL) produced the most consistent improvements in minority class performance while preserving the original dataset. Under imbalance ratio (IR) weighting:

- Logistic Regression (5K) improved minority F1-score to 0.389 and minority recall to 0.835, compared to 0.149 and 0.094 under standard training.
- Random Forest produced more balanced trade offs, achieving moderate F1-scores with better recall than standard training.
- TabNet maintained minority recall above 0.40 across multiple settings.
- Stacking achieved very strong minority recall in smaller datasets, reaching up to 0.894 under IR weighting.
- Some weighting schemes, particularly ICF and LICF, occasionally produced unstable behavior (e.g., extremely high recall with poor specificity).

Overall, CSL consistently outperformed SMOTE across dataset sizes by improving detection of high risk cases without modifying the original dataset.

Table 4: Model performance comparison for 5K sample size and original dataset with IR weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
5K Sample Size									
Test Accuracy	0.777	0.860	0.871	0.871	0.563	0.828	0.8213	0.8853	0.7010
AUC-ROC	0.8541	0.8314	0.7593	0.8293	0.7141	0.7903	0.8015	0.7714	0.8300
F1-Score (L0)	0.8636	0.9221	0.9296	0.9288	0.6951	0.9015	0.8948	0.9383	0.8070
F1-Score (L1)[1]	0.3890	0.3069	0.2367	0.3175	0.2293	0.3246	0.4071	0.1887	0.3370
Recall (L0)	0.7716	0.9060	0.9301	0.9191	0.5443	0.8767	0.8470	0.9718	0.6831
Recall (L1)[1]	0.8353	0.3647	0.2353	0.3529	0.3529	0.4026	0.5974	0.1299	0.8941
Original Dataset									
Test Accuracy	0.7514	0.7532	0.7801	0.7413	0.7247	0.7089	0.91	0.9146	0.7514
AUC-ROC	0.8439	0.8386	0.8374	0.8473	0.7850	0.8442	0.8298	0.8486	0.8447
F1-Score (L0)	0.8483	0.8453	0.8550	0.8375	0.8312	0.8355	0.95	0.955	0.8456
F1-Score (L1)[1]	0.3642	0.3592	0.3630	0.3577	0.3192	0.3352	0.16	0.1687	0.3622
Recall (L0)	0.7525	0.7481	0.7652	0.7343	0.7318	0.8355	0.99	0.9931	0.7480
Recall (L1)[1]	0.7815	0.7781	0.7496	0.8044	0.7064	0.6967	0.09	0.0987	0.7863

Table 5: Model performance comparison for 5K sample size and original dataset with Custom weighting (1/IR)

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
5K Sample Size									
Test Accuracy	0.7638	0.9435	0.9973	0.9428	0.5535	0.828	0.7973	0.8853	0.0850
AUC-ROC	0.8650	0.8307	0.7593	0.8323	0.7141	0.7903	0.8002	0.7714	0.8277
F1-Score (L0)	0.8591	0.9223	0.9296	0.9550	0.6951	0.9015	0.8784	0.9383	0.0000
F1-Score (L1)[1]	0.3893	0.2930	0.2367	0.0652	0.2293	0.3246	0.392	0.1887	0.1567
Recall (L0)	0.7628	0.9082	0.9300	0.9957	0.5443	0.8767	0.8158	0.9718	0.0000
Recall (L1)[1]	0.8588	0.3412	0.2353	0.0353	0.7647	0.4026	0.6363	0.1299	1.0000
Original Dataset									
Test Accuracy	0.7549	0.7515	0.9102	0.9102	0.7295	0.7089	0.91	0.9146	0.7533
AUC-ROC	0.8439	0.8390	0.8446	0.8459	0.7850	0.8442	0.8303	0.8486	0.8449
F1-Score (L0)	0.8482	0.8458	0.9530	0.9530	0.8312	0.8136	0.95	0.955	0.8470
F1-Score (L1)[1]	0.3640	0.3604	0.0009	0.0005	0.3192	0.3352	0.15	0.1687	0.3627
Recall (L0)	0.7523	0.7487	1.0000	1.0000	0.7317	0.6967	0.99	0.9931	0.7504
Recall (L1)[1]	0.7813	0.7799	0.0005	0.0002	0.7064	0.8355	0.09	0.0987	0.7822

[1]Minority class (Level 1: heart condition) metric.

Table 6: Model performance comparison for 5K sample size and original dataset with ICF weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
5K Sample Size									
Test Accuracy	0.3920	0.4920	0.8330	0.7730	0.5170	0.680	0.1027	0.8853	0.0850
AUC-ROC	0.8642	0.7966	0.7194	0.8090	0.7146	0.8086	0.7901	0.7714	0.7998
F1-Score (L0)	0.5033	0.6175	0.9065	0.8646	0.6508	0.7895	0.0000	0.9383	0.0000
F1-Score (L1)[1]	0.2165	0.2440	0.2160	0.2972	0.2172	0.3330	0.1862	0.1887	0.1567
Recall (L0)	0.3366	0.4481	0.8852	0.7924	0.4918	0.6687	0.0000	0.9718	0.0000
Recall (L1)[1]	0.9882	0.9647	0.2706	0.5647	0.7882	0.7792	1.0000	0.1299	1.0000
Original Dataset									
Test Accuracy	0.3266	0.2758	0.4833	0.4231	0.7295	0.3338	0.91	0.9146	0.0898
AUC-ROC	0.8432	0.8234	0.8224	0.8440	0.7850	0.8426	0.8322	0.8486	0.8372
F1-Score (L0)	0.4142	0.3403	0.6069	0.5380	0.8312	0.4258	0.95	0.955	0.0000
F1-Score (L1)[1]	0.2084	0.1973	0.2467	0.2323	0.3192	0.2066	0.15	0.1687	0.1648
Recall (L0)	0.2615	0.2052	0.4381	0.3690	0.7317	0.2708	0.99	0.9931	0.0000
Recall (L1)[1]	0.9873	0.9914	0.9423	0.9722	0.7064	0.9877	0.09	0.0987	1.0000

Table 7: Model performance comparison for 5K sample size and original dataset with LICF weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
5K Sample Size									
Test Accuracy	0.6190	0.7590	0.8560	0.8250	0.5390	0.776	0.6053	0.8853	0.0850
AUC-ROC	0.8655	0.8148	0.7573	0.8130	0.7141	0.7837	0.7967	0.7714	0.8205
F1-Score (L0)	0.7392	0.8519	0.9204	0.8995	0.6724	0.8679	0.7234	0.9383	0.0000
F1-Score (L1)[1]	0.2931	0.3539	0.2421	0.3243	0.2226	0.2632	0.3116	0.1887	0.1567
Recall (L0)	0.5902	0.7574	0.9104	0.8557	0.5169	0.8202	0.5750	0.9718	0.0000
Recall (L1)[1]	0.9294	0.7765	0.2706	0.4941	0.7765	0.3896	0.8701	0.1299	1.0000
Original Dataset									
Test Accuracy	0.5908	0.5439	0.6392	0.5972	0.7295	0.5887	0.91	0.9146	0.0898
AUC-ROC	0.8438	0.8357	0.8336	0.8468	0.7850	0.8467	0.8303	0.8486	0.8436
F1-Score (L0)	0.7128	0.6683	0.7564	0.7186	0.8312	0.7114	0.95	0.955	0.0000
F1-Score (L1)[1]	0.2883	0.2705	0.3046	0.2910	0.3192	0.2845	0.16	0.1687	0.1648
Recall (L0)	0.5579	0.5047	0.6154	0.5652	0.7317	0.5558	0.99	0.9931	0.0000
Recall (L1)[1]	0.9235	0.9420	0.8805	0.9210	0.7064	0.9309	0.09	0.0987	1.0000

[1]Minority class (Level 1: heart condition) metric.

Best models and Hyperparameter Optimization

Across all sample sizes, Logistic Regression tended to be most effective in smaller datasets (5K–20K) under CSL, while Random Forest and TabNet became stronger at larger scales (50K and above). Stacking consistently emerged as the strongest hybrid approach, indicating that combining heterogeneous learners improves generalization under imbalance. This supports the main contribution of the study: model suitability is sample size dependent, and choosing one "best" model without considering dataset scale can reduce clinical sensitivity.

Table 8: Best Model Performance Comparison for Different Dataset Sizes

Dataset Size	Best ML Model	Best DL Model	Best Hybrid Model
5K	Logistic Regression (Custom Weight)	MLP (IR)	Stacking Model (IR)
10K	Logistic Regression (Custom Weight)	MLP (IR)	Stacking Model (Custom Weight)
20K	Logistic Regression (Custom Weight)	MLP (Custom Weight)	Stacking Model (Custom Weight)
50K	Random Forest (LICF)	MLP (ICF)	Stacking Model (ICF)
100K	Random Forest (Custom Weight)	TabNet (IR)	Stacking Model (Custom Weight)
Original Dataset	Random Forest (Custom Weight)	TabNet (Custom Weight)	Stacking Model (Custom Weight)

IR = Imbalance Ratio; ICF = Inverse Class Frequency; LICF = Log of Inverse Class Frequency.

Hyperparameter optimization (Optuna)

Optuna tuning further improved the best models. On the full dataset, a tuned Random Forest (25 estimators, depth 20) achieved minority F1 = 0.39 and AUC–ROC = 0.856. On the 5K subset, Logistic Regression with elastic net regularization achieved minority F1 = 0.38 and AUC–ROC = 0.859. These results indicate CSL enhanced models remain adaptable across dataset sizes and benefit from controlled regularization and depth constraints..

Table 9: Optimal Hyperparameters and Performance Metrics for Best Models

Best Hyperparameters	Performance Metrics
5K Sample Size	
Logistic Regression (Best ML Model)	
penalty = elasticnet; solver = saga;	
C = 0.0326; l1_ratio = 0.1479; max_iter = 68	F1-Score (Level 1): 0.38
AUC-ROC: 0.859	
Accuracy: 0.76	
Precision: 0.40	
Recall: 0.35	
TabNet (Best DL Model)	
cat_emb_dim = 10; learning_rate = 0.078;	
gamma = 0.582; step_size = 81;	
mask_type = sparsemax; batch_size = 256	F1-Score (Level 1): 0.37
AUC-ROC: 0.737	
Accuracy: 0.81	
Precision: 0.42	
Recall: 0.33	
Original Dataset	
Random Forest (Best ML Model)	
n_estimators = 25; max_depth = 20;	
min_samples_split = 4; min_samples_leaf = 6	F1-Score (Level 1): 0.39
AUC-ROC: 0.856	
Accuracy: 0.82	
Precision: 0.43	
Recall: 0.36	
TabNet (Best DL Model)	
cat_emb_dim = 4; learning_rate = 0.0195;	
gamma = 0.863; step_size = 28;	
mask_type = entmax; batch_size = 128	F1-Score (Level 1): 0.34
AUC-ROC: 0.845	
Accuracy: 0.72	
Precision: 0.39	
Recall: 0.31	

DL = Deep Learning; ML = Machine Learning.

Explainable AI (SHAP) – Clinical Interpretation

SHAP analysis identified AgeCategory, ChestScan, and GeneralHealth as dominant predictors. This is clinically consistent: age is a well known non modifiable risk factor; chest imaging often correlates with higher comorbidity bur-

den or prior investigation pathways more common in high risk patients; and self reported general health captures latent factors such as chronic disease burden, frailty, and functional limitation. Variables such as smoking and BMI appeared lower in SHAP rankings, which can occur in survey derived data where (i) their effects are mediated through downstream factors (physical activity, general health, mobility difficulty) and (ii) risk may manifest indirectly through comorbidities.

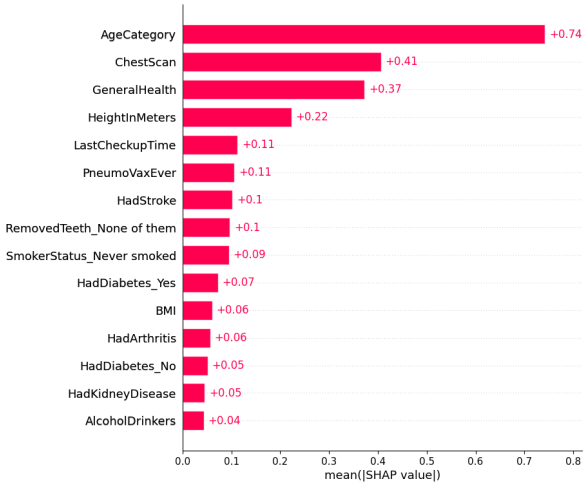

Fig. 5: Mean absolute SHAP values ranking features by their overall importance in predicting heart condition, indicating age category, chest scans, and general health as the most influential variables.

The SHAP summary also showed that higher age categories push predictions toward Level 1, matching both EDA trends and clinical expectation.

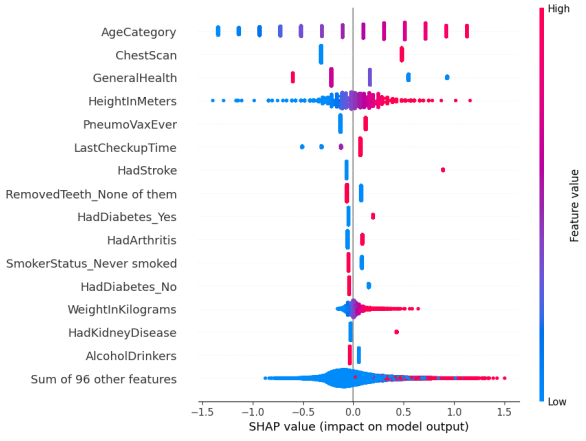

Fig. 6: SHAP summary plot illustrating the contribution and distribution of top features influencing heart condition prediction, with color intensity representing feature values from low (blue) to high (red).

Force plots enabled case level interpretability by showing how each feature moved the prediction toward or away from a heart condition classification.

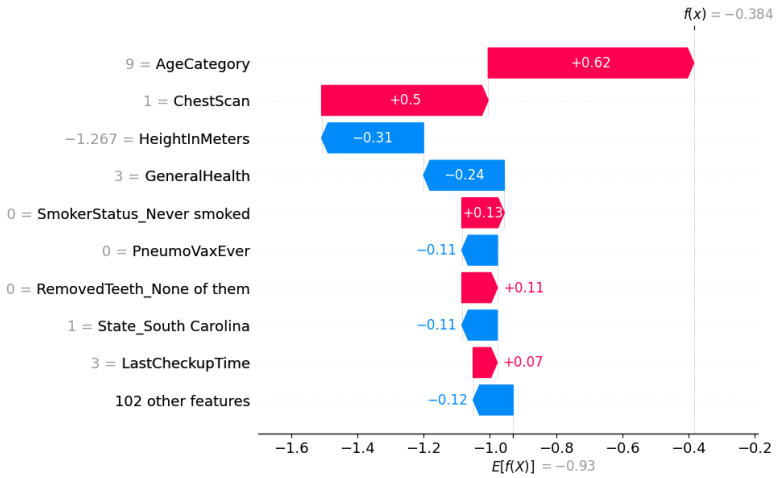

Fig. 7: SHAP force plot for the 5th observation in the test set, showing individual feature contributions to the model's prediction. Positive (red) and negative (blue) forces indicate features pushing the prediction toward or away from a heart condition.

Feature Contribution and Inclusion Analysis

Feature inclusion was guided by statistical association measures obtained during exploratory data analysis and by model based importance techniques.
Cramér's V analysis identified *GeneralHealth*, *AgeCategory*, *ChestScan*, *DifficultyWalking*, and *RemovedTeeth* as the most strongly associated categorical predictors of heart conditions. Correlation analysis revealed multicollinearity between *WeightInKilograms* and *BMI*, confirming redundancy between these numerical features and informing interpretation during modeling.
SHAP analysis further validated the influence of key predictors, with *AgeCategory*, *GeneralHealth*, and *ChestScan* consistently emerging as dominant contributors to model predictions. The consistency between statistical association measures (Chi square and Cramér's V), tree based importance (Random Forest), and SHAP attribution strengthened the reliability of feature interpretation.
In the RF + TabNet hybrid model, Random Forest importance ranking was used to retain the top 85% of features (38 → 32 features), reducing dimensionality while preserving predictive performance. This controlled feature reduction ensured computational efficiency while maintaining clinically relevant predictors.
Overall, feature handling was not based on arbitrary removal but on statistically supported and model validated importance measures, ensuring methodological consistency and medical interpretability.

Quantum Machine Learning: proof of concept, not competitive benchmarking

Quantum Support Vector Machine (QSVM) was implemented as an exploratory proof of concept rather than as a competitive benchmarking model.
Due to computational limitations, experiments were restricted to:

- 500 - 1100 samples,
- Three selected features.

Under these constraints, classical kernels (polynomial and RBF) generally outperformed QSVM. However, QSVM performance improved with increasing sample size and achieved competitive results at 1,100 samples relative to sigmoid kernels.
Given the limited feature space and hardware constraints, QSVM results should be interpreted as exploratory evidence of feasibility rather than performance superiority. Future research utilizing scalable quantum feature maps and larger qubit resources may yield stronger results.

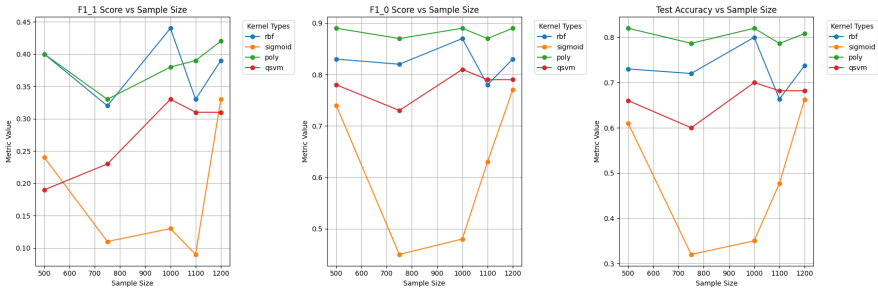

Fig. 8: SVM performance across different kernels and sample sizes based on F1 scores and test accuracy.

Cost sensitive learning emerged as the most reliable imbalance-handling strategy. Logistic Regression and Random Forest were the strongest ML models depending on dataset size. TabNet demonstrated robust DL performance for larger datasets. Stacking provided consistent hybrid improvements.SHAP ensured interpretability aligned with established cardiovascular risk structure. QSVM demonstrated exploratory potential but remains proof of concept under current computational limits. Together, these findings provide a structured, scalable framework for heart condition prediction under severe class imbalance while maintaining interpretability and methodological clarity.

4.3 Discussion

The findings of this study provide structured methodological insights into the application of machine learning (ML), deep learning (DL), hybrid, and quantum machine learning (QML) models for heart condition prediction under

severe class imbalance. The results confirm that class imbalance remains a critical challenge in medical prediction tasks. As reported in prior studies, accuracy alone becomes misleading in imbalanced settings, since models may favor the majority class while failing to detect clinically critical minority cases (Akosa, 2017). This behavior was clearly observed in the present study, where standard training produced high overall accuracy (above 0.91) but extremely low recall for heart condition cases.

The Synthetic Minority Oversampling Technique (SMOTE), although widely adopted in imbalance learning, did not consistently improve minority-class performance in this dataset. While recall improved in certain models, instability and reduced overall generalization were observed. In medical contexts, synthetic interpolation between minority samples may distort underlying physiological relationships, especially when categorical and clinically interdependent features are present. These findings suggest that data level balancing approaches must be applied cautiously in healthcare analytics.

In contrast, cost sensitive learning (CSL) demonstrated consistent and clinically meaningful improvements across ML, DL, and hybrid models. By assigning higher misclassification costs to heart condition cases, CSL directly optimized the decision boundary without modifying the original data distribution. This strategy aligns with recent medical imbalance research emphasizing the importance of preserving authentic patient data while adjusting learning priorities (Araf et al., 2024). Among the tested weighting schemes, imbalance ratio (IR) weighting provided the most stable trade off between recall and precision, suggesting its suitability for large scale epidemiological datasets.

When comparing model categories, traditional ML algorithms particularly Logistic Regression and Random Forest exhibited strong and stable performance across varying dataset sizes. Logistic Regression performed competitively in smaller samples, while Random Forest demonstrated greater robustness as dataset size increased. These findings support prior evidence that classical statistical learning models often remain highly effective for structured clinical datasets (Latifah et al., 2020).

Among DL approaches, the Multilayer Perceptron (MLP) showed better adaptability in smaller datasets under CSL, whereas TabNet demonstrated improved scalability for larger datasets due to its sequential attention based feature selection mechanism. Hybrid strategies, particularly stacking, consistently achieved the highest balanced performance, indicating that combining diverse decision boundaries enhances generalization in heterogeneous medical data.

From an interpretability perspective, SHAP based analysis provided clinically coherent explanations. Age category emerged as the most influential predictor, which aligns with established cardiovascular risk evidence indicating increasing risk after midlife (Rodgers et al., 2019). General health status, smoking history, BMI, and chest scan history also showed strong contributions, all of which are recognized cardiovascular risk factors in epidemiological research. Importantly, SHAP beeswarm and force plots demonstrated that higher age groups and poorer health ratings consistently pushed predictions

toward positive classification, reinforcing medical plausibility. This alignment between model explanations and known risk factors enhances trustworthiness and clinical interpretability.

The Quantum Machine Learning (QML) component was implemented explicitly as a proof of concept rather than a competitive benchmarking exercise. Due to qubit limitations and computational constraints, the Quantum Support Vector Machine (QSVM) was evaluated using reduced feature subsets and small sample sizes. While classical kernels generally outperformed QSVM under current hardware conditions, the observed improvements with increasing sample size suggest potential scalability in future quantum enhanced healthcare applications. Therefore, QML should be interpreted as exploratory groundwork rather than a replacement for classical ML at this stage.

Overall, this study demonstrates that algorithm level imbalance handling through cost sensitive learning is more reliable than synthetic oversampling for heart condition prediction using large epidemiological datasets. Classical ML models remain highly competitive, hybrid strategies enhance stability, and explainable AI ensures clinical transparency. The integration of CSL with interpretable modeling provides a balanced framework for ethically grounded and practically deployable cardiac risk prediction systems.

Future research should investigate feature interaction stability under different imbalance ratios, explore clinically guided cost matrix calibration, and evaluate quantum enhanced kernels under expanded hardware capacity. Expanding dataset diversity beyond BRFSS and incorporating multi-center clinical validation would further strengthen the translational applicability of predictive models in cardiovascular healthcare.

5 Conclusion

This study systematically evaluated machine learning (ML), deep learning (DL), hybrid, and quantum machine learning (QML) approaches for heart condition prediction under severe class imbalance. The primary objective identifying the most reliable modeling strategy across varying dataset sizes while preserving medical data integrity was achieved.

The findings demonstrate that algorithm level imbalance handling through cost sensitive learning (CSL) is more effective than synthetic oversampling methods such as SMOTE. CSL consistently improved minority class recall, F1 score, and AUC-ROC without altering the original data distribution, making it more suitable for large scale epidemiological healthcare datasets.

Among classical ML models, Logistic Regression and Random Forest showed stable and competitive performance across dataset sizes, with Logistic Regression performing strongly in smaller samples and Random Forest demonstrating robustness at larger scales. In the DL category, MLP exhibited better adaptability in smaller datasets, whereas TabNet showed superior scalability for larger datasets due to its sequential attention-based feature selection. Hybrid strategies, particularly stacking, consistently delivered the most balanced

and robust results, highlighting the benefit of integrating complementary decision boundaries.

Explainable AI (XAI) through SHAP analysis ensured clinical interpretability, with identified key predictors such as age, general health status, BMI, and smoking history aligning with established cardiovascular risk factors. This alignment strengthens the practical reliability of the proposed framework.

The QML component, implemented as a proof of concept using QSVM, demonstrated feasibility but remained constrained by current quantum hardware limitations and reduced feature space. While classical models outperformed QSVM under present computational conditions, the results suggest potential for future scalability as quantum resources mature.

Overall, this study provides a structured and interpretable modeling framework for heart condition prediction, emphasizing cost-sensitive learning as a reliable imbalance handling strategy while integrating explainability and exploratory quantum analysis. These findings contribute toward developing accurate, ethical, and clinically deployable predictive systems in cardiovascular healthcare.

Code Repositories

Available at: `https://github.com/Maleesha-98/Final-Year-Project--ST-4057/tree/main/Codes`

Appendices

Detailed experimental results and extended tables are available at:

[6pt]

References

Adhikary D., Barman S., Ranjan R. and Stone H. (2022) 'A systematic review of major cardiovascular risk factors: A growing global health concern', *Cureus*, 14(10): e30119. DOI: 10.7759/cureus.30119.

Ahmad A.A. and Polat H. (2023) 'Prediction of heart disease based on machine learning using jellyfish optimization algorithm', *Diagnostics*, 13(14): 2392.

Akosa J. (2017) 'Predictive accuracy: A misleading performance measure for highly imbalanced data', *Proceedings of the SAS Global Forum*, 12: 1–4.

Albert A.J., Murugan R. and Sripriya T. (2023) 'Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology', *Research on Biomedical Engineering*, 39(1): 99–113.

Al'Aref S.J., Anchouche K., Singh G., Slomka P.J., Kolli K.K., Kumar A., Pandey M., Maliakal G., Van Rosendael A.R. and Beecy A.N. (2019)

'Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging', *European Heart Journal*, 40(24): 1975–1986.

Araf I., Idri A. and Chairi I. (2024) 'Cost-sensitive learning for imbalanced medical data: a review', *Artificial Intelligence Review*, 57(4): 80.

Athanasiou M., Sfrintzeri K., Zarkogianni K., Thanopoulou A.C. and Nikita K.S. (2020) 'An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus', *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*: 859–864.

Babu S.V., Ramya P. and Gracewell J. (2024) 'Revolutionizing heart disease prediction with quantum-enhanced machine learning', *Scientific Reports*, 14(1): 7453.

Benhar H., Idri A. and Fernández-Alemán J.L. (2020) 'Data preprocessing for heart disease classification: A systematic literature review', *Computer Methods and Programs in Biomedicine*, 195: 105635.

Bhatt C.M., Patel P., Ghetia T. and Mazzeo P.L. (2023) 'Effective heart disease prediction using machine learning techniques', *Algorithms*, 16(2): 88.

Buchan T.A., Ross H.J., McDonald M., Billia F., Delgado D., Posada J.G.D., Luk A., Guyatt G.H. and Alba A.C. (2019) 'Physician prediction versus model predicted prognosis in ambulatory patients with heart failure', *The Journal of Heart and Lung Transplantation*, 38(4): S381.

Chicco D. and Jurman G. (2020) 'Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone', *BMC Medical Informatics and Decision Making*, 20(1): 16.

D'Agostino Sr R.B., Vasan R.S., Pencina M.J., Wolf P.A., Cobain M., Massaro J.M. and Kannel W.B. (2008) 'General cardiovascular risk profile for use in primary care: the Framingham Heart Study', *Circulation*, 117(6): 743–753.

Elkan C. (2001) 'The foundations of cost-sensitive learning', *International Joint Conference on Artificial Intelligence*, 17(1): 973–978.

Fernández A., Garcia S., Herrera F. and Chawla N.V. (2018) 'SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary', *Journal of Artificial Intelligence Research*, 61: 863–905.

Garavand A., Salehnasab C., Behmanesh A., Aslani N., Zadeh A.H. and Ghaderzadeh M. (2022) 'Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms', *Journal of Healthcare Engineering*, 2022(1): 5359540.

Hajar R. (2017) 'Risk factors for coronary artery disease: historical perspectives', *Heart Views*, 18(3): 109–114.

Javaheri S., Omobomi O. and Redline S. (2019) 'Insufficient sleep and cardiovascular disease risk', in *Sleep and Health*, 203–212.

Latifah F.A., Slamet I. and Sugiyanto S. (2020) 'Comparison of heart disease classification with logistic regression algorithm and random forest algorithm', *AIP Conference Proceedings*, 2296.

Loh D.R., Yeo S.Y., Tan R.S., Gao F. and Koh A.S. (2022) 'Explainable machine learning predictions to support personalized cardiology strategies', *European Heart Journal – Digital Health*, 3(1): 49–55.

Mahmood S.S., Levy D., Vasan R.S. and Wang T.J. (2014) 'The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective', *The Lancet*, 383(9921): 999–1008.

Natarajan K., Vinoth Kumar V., Mahesh T.R., Abbas M., Kathamuthu N., Mohan E. and Annand J.R. (2024) 'Efficient heart disease classification through stacked ensemble with optimized firefly feature selection', *International Journal of Computational Intelligence Systems*, 17(1): 174.

Pillai A.S. (2022) 'Cardiac Disease Prediction with Tabular Neural Network', *International Journal of Engineering Research and Technology*, 11(11): 99–102. DOI: 10.17577/IJERTV11IS110099.

Rodgers J.L., Jones J., Bolleddu S.I., Vanthenapalli S., Rodgers L.E., Shah K., Karia K. and Panguluri S.K. (2019) 'Cardiovascular risks associated with gender and aging', *Journal of Cardiovascular Development and Disease*, 6(2): 19.

Sumida K., Molnar M.Z., Potukuchi P.K., Thomas F., Lu J.L., Yamagata K., Kalantar-Zadeh K. and Kovesdy C.P. (2019) 'Constipation and risk of death and cardiovascular events', *Atherosclerosis*, 281: 114–120.

Tjoa E. and Guan C. (2021) 'A survey on explainable artificial intelligence (XAI): Toward medical XAI', *IEEE Transactions on Neural Networks and Learning Systems*, 32(11): 4793–4813. DOI: 10.1109/TNNLS.2020.3027314.

Tompra K.-V., Papageorgiou G. and Tjortjis C. (2024) 'Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification', *Algorithms*, 17(5): 178.

Truett J., Cornfield J. and Kannel W. (1967) ‘A multivariate analysis of the risk of coronary heart disease in Framingham’, *Journal of Chronic Diseases*, 20(7): 511–524.

Wilson P.W.F., D’Agostino R.B., Levy D., Belanger A.M., Silbershatz H. and Kannel W.B. (1998) ‘Prediction of coronary heart disease using risk factor categories’, *Circulation*, 97(18): 1837–1847.

World Health Organization (2025) ‘Cardiovascular diseases’. Available at: https://www.who.int/health-topics/cardiovascular-diseases (Accessed: 27 January 2025).

World Heart Federation (2023a) ‘Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report’. Available at: https://world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/ (Accessed: 28 January 2025).

World Heart Federation (2023b) ‘World Heart Report 2023: Confronting the World’s Number One Killer’. Available at: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf (Accessed: 27 January 2025).