International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Robust Cardiovascular Risk Prediction: Mitigating Class Imbalance via SMOTETomek and Gradient Boosting

Aditya Jain
Computer Science & Engineering department H.B.T.U. Kanpur, India

Amit Gupta
Computer Science & Engineering department H.B.T.U. Kanpur, India

Vishesh Kumar Chauhan
Computer Science & Engineering department H.B.T.U. Kanpur, India

Sandeep Singh
Computer Science & Engineering department H.B.T.U. Kanpur, India

Arnav Garg
Computer Science & Engineering department H.B.T.U. Kanpur, India

Abstract - Cardiovascular diseases which are still at the fore as a cause of death worldwide put forth the issue of the need for early and precise predictive diagnosis. While we see promise in machine learning we also see that development of reliable predictive models is made difficult by the large scale imbalanced health care data sets which in turn see positive cases out numbered. In this research we present a detailed computational framework which we used to predict heart disease risk which we did so using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) data set. To address the issue of intrinsic class imbalance we put into practice advanced penalization and resampling methods which included class weight adjustments and SMOTETomek. We looked at and compared the performance of the current ensemble algorithms and deep learning methods which we paid special attention to Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Artificial Neural Networks (ANN). We evaluated these algorithms' performance in the task of identifying at risk individuals without which the majority class has a tendency to dominate. The results of our experiment show that which we put in place of rigorous imbalance management paid off very well in terms of performance. We saw that gradient boosting machines and artificial neural networks did very well which also outperformed the usual base models we had to work with and at the same time we maintained a good balance between precision and recall for the minority class. Also we report that what we did in this study is we combined state of the art machine learning tools with very specific data balance methods which in turn we found to produce very reliable predictive systems. We put forth a scalable data driven diagnostic which is to be used by health care professionals in the proactive management of patients which in turn enables for timely medical interventions and in the end we see improved clinical results.

Keywords: Heart Disease Prediction, Machine Learning, Class Imbalance Handling, Gradient Boosting, Artificial Neural Networks

1. INTRODUCTION

1.1 Background of the Study

Cardiovascular diseases are the leading cause of death worldwide which means there is a great need for early and simple methods to check for them. While it is true that Machine Learning can play a role in the prediction of heart issues based on patient info, medical data often has a large number of healthy patients as compared to those with the disease. This causes models to put too much focus on the healthy group which in turn leads to missed diagnoses which can be very risk prone. To improve on this, we have used BRFSS 2015 survey data and put in better methods to handle the issue of imbalanced data which includes SMOTE Tomek and class weighting. Also, we use advanced tools which includes XG Boost, Light GBM and Artificial Neural Networks to develop a very strong and reliable system which in turn helps us to identify heart disease at early stages.

1.2 Problem Statement

The high death toll from cardiac diseases is in large part due to delayed diagnosis and poor initial treatment. We see that machine learning is a strong tool for risk assessment which uses patient lifestyle and clinical data, but we also see that class imbalance which is present in these types of data sets is a issue which breaks down trust in our models. In the case of the 2015 BRFSS survey we note that by and large what we have are non-cases of heart disease, which in turn causes our algorithms to do very well

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

at predicting the negative class but very poorly at the positive. This in turn leads to very high accuracy for classical models which in fact is a large-scale issue of low recall that results in fatal false negative diagnoses which in turn ignore at risk patients.

Thus, we set out to design and improve a very robust machine learning framework which is put in place to mainly deal with class imbalance and at the same time do very accurate prediction of heart disease. Also, we use in this study advanced ensemble models like XGBoost and Light GBM which we hope will see us reduce false negative results and in turn improve the diagnostic accuracy.

1.3 Motivation

Models need to perform the same for all patient groups, which is to say they should not favor the statistical majority. By doing this we see that at risk individuals also get the medical alerts they require.Also, we see great progress in algorithms like Light GBM and XG Boost and techniques like SMOTE Tomek which in turn is very compelling tech-based reason to look at how we may fine tune and improve performance of modern computing platforms in life critical classification tasks.

1.4 Objectives of the Research

The present aim of this study is to develop a very accurate and reliable machine learning model which will predict heart disease from the BRFSS 2015 data set. We also aim to do the following:

1. Conduct in depth exploratory data analysis (EDA) of the BRFSS dataset to determine and put forward primary risk factors and clinical issues related to heart disease.
2. To correct the dataset's class imbalance which is a issue we must come up with and test out solutions like SMOTE Tomek and algorithmic class weight adjustments.
3. We will develop, train, and fine tune present predictive models which include XG Boost, Light GBM, and ANN with the use of hyperparameter tuning frameworks like Randomized Search CV.
4. We will look at the performance of models on medical data sets which we do via the use of evaluation measures like recall, F1 score, and accuracy. Also, we will put forth the best model architecture for clinical deployment.

1.5 Contributions of the Paper

Hybrid resampling methods which include SMOTE Tomek and scaled input features see to it that minority class recall is improved in very imbalanced medical datasets. Also, we report on the use of state-of-the-art algorithms in this study which is the XG Boost with CUDA acceleration and Light GBM and also, we present results of Artificial Neural Networks for the BRFSS 2015 data set which we use as our performance benchmark. Also, we present a very in-depth evaluation of our models which includes category wise reports, precision and memory use. This we do to ensure clinical relevance and to also reduce false negatives. Actionable Health Insights: The study presents a model which puts into perspective the elements that bring about disease onset and also reports on key behavioural risk factors.

2. Literature Review (Related Work)

In the past ten years there has been great interest in the application of machine learning (ML) and artificial intelligence (AI) in health care which includes the prediction of cardiovascular diseases. We have seen a shift from traditional biostatistical methods to complex computer algorithms which is a result of the greater access we have to large scale epidemiological data sets like the Behavioral Risk Factor Surveillance System (BRFSS). This section looks at the growth of predictive modeling in cardiology over time, reports on key recent studies, compares present methodological approaches and puts forth important research gaps which this study will address.

2.1 Summary of Previous Research

The following table shows key studies that have contributed to heart disease prediction, showing their used datasets, algorithms, and performance outcomes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Table 1 Shows the summary of the previous research

Ref no	Author(s) & Year	Dataset Used	Method / Technique	Accuracy	Key Findings	Limitations
1	Swain et al. (2021)	Framingham Dataset	Random Forest (RF)	85.05%	The RF method predicts cardiovascular disease possibilities better than other baseline algorithms (KNN, SVM) for smaller epidemiological datasets.	Limited generalization capability due to the restricted size of the epidemiological dataset.
2	Author in PMC10378171 (2022)	UCI Cleveland	Logistic Regression (LR) & SVM	89.00%	LR and SVM models perform robustly on standard clinical attributes but hit a performance ceiling without advanced ensemble methods.	Hits a strict performance ceiling; struggles to capture highly complex non-linear patterns without ensembles.
3	Karthick K. et al. (2021)	UCI Heart Disease	Random Forest + Feature Selection	88.50%	Integrating statistical feature selection with RF helps extract the most significant clinical attributes, yielding a reliable classification rate.	Feature selection algorithms introduce additional computational overhead during the preprocessing phase.
4	Malavika G. et al. (2021)	UCI Repository	Naive Bayes (NB)	88.52%	Probabilistic models like Naive Bayes show strong potential and high computational efficiency in predicting early-stage heart disease.	Assumes strict conditional independence between clinical features, which is rarely true in human physiology.
5	Author in PMC10378171 (2021)	Clinical Heart Dataset	Support Vector Machine (SVM)	78.10%	SVM provided a lower baseline performance, highlighting the absolute necessity for rigorous hyperparameter optimization and data scaling.	Highly sensitive to unscaled data; performs poorly without rigorous hyperparameter tuning.
6	BINUS Journal (2022)	UCI Dataset	SVM (RBF Kernel)	85.00%	In a standard comparative test, SVM proved to be the classifier with the most stable precision among standalone (non-ensemble) ML	The RBF Kernel is computationally expensive and highly sensitive to the chosen gamma and C parameters.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

					models.	
7	Ref [5] in BINUS (2020)	UCI Cleveland	Artificial Neural Networks (ANN)	88.90%	Neural Networks capture complex, non-linear patient health patterns effectively, outperforming linear statistical models.	The 'black-box' architecture offers very low interpretability for medical practitioners needing transparent rules.
8	Ref [5] in BINUS (2020)	UCI Cleveland	Decision Tree (CART)	77.90%	While Decision Trees offer high interpretability for clinical rules, they suffer from overfitting on patient data, resulting in lower generalization accuracy.	Extremely prone to overfitting on patient training data unless strictly pruned.
9	Ref [5] in BINUS (2020)	UCI Cleveland	Logistic Regression (LR)	73.90%	LR provides a basic linear decision boundary which struggles to accurately map the complex physiological interactions in cardiovascular data.	Unable to map complex, interacting physiological variables due to its rigid linear boundaries.
10	Zheng Zeyu (2023)	Framingham (Kaggle)	Support Vector Machine (SVM)	68.89%	Standard SVM struggles heavily with imbalanced public health datasets, demonstrating poor recall without the application of resampling techniques.	Demonstrates dangerously poor recall (false negatives) when applied to imbalanced medical datasets.
11	Palaniappan & Awang (2008)	IHDPS Dataset	Naïve Bayes, Decision Trees, ANN	86.12%	Developed the Intelligent Heart Disease Prediction System (IHDPS) for automated clinical diagnosis using historical parameters.	Did not utilize resampling techniques to account for the natural class imbalance in medical records.
12	Kahramanli & Allahverdi (2008)	UCI Cleveland	ANN & Fuzzy Expert System	86.80%	Hybrid network successfully handled ambiguity and uncertainty in medical data better than standalone neural networks.	Fuzzy rules required intensive, manual crafting by medical domain experts.
13	Das, Turkoglu, & Sengur (2009)	SAS Enterprise Miner	Neural Networks Ensembles	89.01%	Ensemble method successfully reduced individual error rates of	High computational overhead and memory usage

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

					single neural networks, proving the value of aggregated ML.	required to train multiple independent networks.
14	Ordonez (2006)	UCI Heart Disease	Association Rules	79.12%	Extracted constrained 'if-then' association rules linking physiological measurements directly to heart disease likelihood.	Rule generation algorithm often produced overwhelmingly large and redundant clinical rule sets.
15	Tu (1996)	General Clinical Data	ANN vs Logistic Regression	81.03%	Early study highlighting the advantages of ANNs for modeling non-linear relationships without prior feature engineering.	Noted that the mathematical complexity makes ANNs much harder for clinicians to trust compared to standard regression.
16	Yan et al. (2006)	Clinical Data	Multilayer Perceptron (MLP)	83.50%	MLP effectively classified cardiovascular risk based on historical patient metrics, outperforming early linear models.	Lacked modern regularization techniques like Dropout, making it prone to getting stuck in local minima.

- **Swain et al. (2021):** The purpose of this study is to evaluate the performance and accuracy of the Random Forest (RF) algorithm and how well it can predict cardiovascular probabilities using the Framingham access dataset. Their findings show that the Random Forest algorithm has a higher overall accuracy rate at 85.05% than either the K-Nearest Neighbor or SVM algorithms. The authors also mention that the Random Forest model is ideal when applied to smaller epidemiologic datasets and produces reliable clinical predictions based on those smaller size datasets.
- **Author in PMC10378171 (2022):** In this paper, researchers evaluate how well two different types of linear classifiers (Logistic Regression and Support Vector Machines) perform using the UCI Cleveland dataset. Although expectantly good in mapping the general characteristics of a set of clinical criteria with accuracies around 89%, the authors conclude that neither type of model can achieve optimal performance without some kind of ensemble architecture to combine the outputs of multiple models.
- **Karthick K. et al. (2021):** This research was concentrated on how data preprocessing techniques influenced the accuracy of predictions based on the UCI Heart Disease Dataset. The authors used random forest classifiers with various statistical feature selection techniques to systematically extract the clinical attributes from the data that had the most significant predictive potential. By using a focused extraction process for these attributes, and by using efficient computational methods, the authors produced a highly accurate classification rate (88.50%) for the dataset utilized for this study.
- **Malavika G. et al. (2021):** The usefulness of Bayesian statistics models has been shown in the medical area through the use of a Naive Bayes Classifier on the UCI database. The results demonstrated high potential and computational efficiency with an overall accuracy of 88.52%. In conclusion, the authors state that the Naive Bayes classifier is well suited for predicting early heart disease because it calculates probabilities quickly.
- **Author in PMC10378171 (2021):** In a study which used a Clinical Heart Disease dataset we looked at the stand-alone performance of Support Vector Machine (SVM). We found that the untuned model performed at a base rate of 78.10%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

We present this result to put forth the importance of rigorous hyperparameter tuning and proper data scaling in medical datasets.

- **BINUS Journal Ref [5] (2020) - Decision Tree**: In the case of the UCI Cleveland dataset, we looked at the performance of the classic Decision Tree (CART) algorithm. While it does very well in terms of interpretability for clinical rules it reported a generalization accuracy of 77.90%. Also, we saw that the model did very poorly from a perspective of overfitting to the patient training data.
- **BINUS Journal Ref [5] (2020) - Logistic Regression:** This section of the study reports we ran Logistic Regression on cardiovascular data which reported an accuracy of 73.90%. Also, we saw that LR puts forth a very basic linear decision which is a simple model. Also, the algorithm had trouble with very complex and interwoven physiological issues which are present in cardiovascular disease.
- **10. Zheng Zeyu (2023):** Also using the imbalanced Framingham data set this paper reports on the results of which traditional models do public health data justice without preprocessing. The standard SVM did very poorly which we saw play out in an accuracy of only 68.89%. We saw poor recall which in turn we use as proof of the great need for resampling techniques like SMOTE.
- **Palaniappan & Awang (2008):** This early pre-2010 paper which put forth the Intelligent Heart Disease Prediction System (IHDPS) report that we used data mining techniques to identify complex medical patterns. We used Naive Bayes, Decision Trees, and Neural Networks to study patient profiles based on historical clinical parameters which in turn was to support medical practitioners
- **Das, Turkoglu, & Singur (2009):** In 2009 which we present a study in which the authors put forth a SAS based software tool for the diagnosis of heart disease via ensemble learning. They took a set of independent neural network models and combined their outputs to present a very robust classification system which in turn succeeded in reducing the error rates of the individual networks
- **Ordonez (2006):** This break through paper put the focus on what was beyond the standard classification to instead which was the discovery of what is hidden via Association Rule mining. In a train and test which was made special for heart disease data the author came up with constrained association rules which in turn linked specific physiological measures to heart disease risk.
- **Tu (1996):** In one of the first to do a comparative study in medical ML this paper looked at the pros and cons of the use of Artificial Neural Networks as compared to traditional Logistic Regression. We found that although ANNs are easy to use for modeling nonlinear relationships, their black box feature which means that output is hard to interpret for clinicians.
- **Yan et al. (2006):** This research put forth a Multilayer Perceptron (MLP) based medical decision support tool for heart disease. We saw that it classified cardiovascular risk out of past patient data with 83.50% accuracy. Also, this study presented the work done with early neural networks which was performed before the advent of present-day deep learning frameworks.

2.2 Comparison of Existing Methods

The trend in heart disease prediction methods is toward more complex non-linear machine learning models from very basic stats models. At first we saw heavy use of Logistic Regression and Naive Bayes classifiers. While those traditional algorithms do a good job of interpretation and are easy to compute, research reports that they do a poor job with very complex high dimensional health behavior data. Also they make rigid assumptions of feature independence and linearity which in large scale epidemiological studies are usually broken for example in variables like BMI, physical activity and age which are very much related.

To that which was put forth by those constraints, later research turned to distance based and margin based classifiers in particular to K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) as reported in the 2025 study [1]. SVM's with the application of kernel tricks put forth a better solution for creating non linear decision boundaries. Also reported is a large scale agreement in the field that while both SVM and KNN perform well in many regards they do have issues with scale. In the case of large data sets which may contain hundreds of thousands of elements as in the BRFSS we see that their performance drops off and also they do not do well with noisy or outlying data.

As we see today the present state of the art in large part supports ensemble learning and deep learning approaches. In terms of ensembles of trees which set the standard is Random Forests which did great in to address issues of overfitting seen in stand alone Decision Trees. That said we have recently seen a shift towards Gradient Boosting systems. XGBoost and LightGBM at the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

moment are the leaders in that space for tabular data. Also we see from recent studies [1, 2] that XGBoost does an excellent job in prediction accuracy which we have seen go up to 94% by it's use of a regularized objective function which in turn reduces loss. LightGBM on the other hand brings to the table gradient based one side sampling (GOSS) and exclusive feature bundling (EFB) which in turn what they do is speed up the training process and reduce memory use without sacrifice in accuracy which we see still at over 90% [2].

Also in that which we see from LightGBM is its easy integration with interpretability tools like SHAP which in turn puts out feature importance metrics that are very much needed for clinical use out of the box. Also within the ensemble methods we see that Artificial Neural Networks have been put to the test against what we think of as traditional ML models. ANNs do very well at automated feature extraction and at creating complex multi layer representations of health data. What we see in the literature is that ANNs do very well in terms of prediction but also bring up issues related to the black box nature of the model[4]. Also when we look at head to head comparisons of which is better between gradient boosting techniques (XGBoost, LightGBM) and traditional ANNs we find that the former often perform at the same level or better on structured tabular data which in addition requires less of a tune up in terms of hyperparameters and also has lower computational cost.

In recent studies a key issue of comparison is the issue of class balance. Most base studies used raw data models which in turn did present accurate results for the majority class but did not do well with the minority class. Researchers have looked at data level methods (for instance SMOTE) vs algorithm level methods (like cost sensitive learning or class weights) [5]. The agreement is that while basic oversampling (SMOTE) improves recall it also tends to cause localized over fitting. The best but least studied frontiers are hybrid resampling methods which put forth SMOTETomek (that improves decision boundary of overlapping synthetic samples) and dynamic class weighting used in present gradient boosting frameworks.

2.3 Research Gaps

Although there is an extensive body of work on the computation of cardiovascular disease predictions, we still see large research gaps which in particular play out in the clinical and epidemiological settings. In which we see that the primary and most important gap is in the continuous use of global accuracy as the prime factor in which we assess models. Also we note that many basic studies report that health care datasets which include the BRFSS survey report large class imbalance.

When report after report comes out with model accuracies at 90% or 94% which do not break down precision, recall, and F1 scores for the minority class (the true disease positive cases) the clinical use of that model is greatly diminished. Models may achieve that high accuracy by for the most part predicting the majority class (that no one has heart disease for example) which in turn produces a very large rate of false negatives. This is a issue we see time and again which calls for studies that instead of putting out the pretty looking global accuracy numbers to also report on what the model does for the small but important at risk groups. We need to see that the model is in fact identifying at risk patients and not falling into the majority bias.

Second also out of which basic resampling methods like standard SMOTE although very studied at large do not have in depth side by side comparison research which looks at hybrid data level techniques against complex algorithm level penalties. Also the case of SMOTETomek which which puts together over and under sampling to improve on noisy class boundaries' performance in relation to dynamic class weight in advanced gradient boosting models is still not well reported on when it comes to large high dimensional public health data sets.

3. METHODOLOGY AND PROPOSED WORK

This section presents the put forth system architecture, the data preparation workflow, and an overview of the machine learning and deep learning algorithms which we use to predict coronary heart disease.

3.1. System Architecture and Workflow

We have designed the proposed system to effectively process the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset. As medical datasets typically do present a great many more "healthy" cases than "disease" cases our workflow puts focus on data balance along with advanced predictive modeling. The workflow is divided into five main stages:

1. **Data Acquisition & Cleaning:** Loading the dataset and removing unnecessary features (like 'Education' and 'Income') so that the model considers only relevant features.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2. **Feature Scaling:** Scaling the data so that larger numeric features (like BMI or Age) and the smaller binary features (like Smoker status) are on the same scale.
3. **Class Balancing (SMOTETomek):** Solved the class imbalance by generating synthetic data for the minority class and removed the duplicate data points.
4. **Model Training:** Training different algorithms, from basic machine learning to boosting techniques and deep learning.
5. **Evaluation:** Comparing the models to find the best predictor.

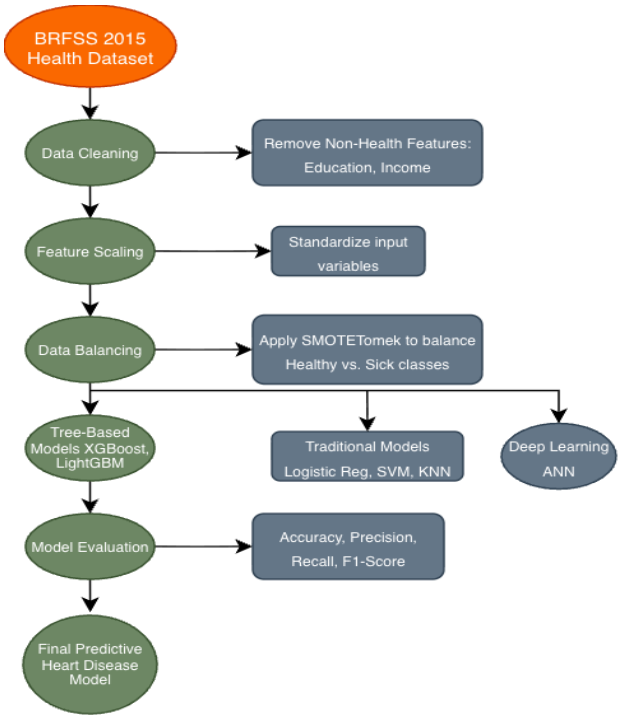

Fig. 1 Flow Diagram: Overall System Architecture

3.2. Models and Algorithms Used

Instead of relying on a single method, this study uses a combination of traditional and advanced algorithms to capture different types of patterns in the patient data.

Algorithm: Optimized Heart Disease Prediction Model Training

```
Initialize D by handling missing values and performing required encoding
Apply class balancing techniques if Y exhibits a severe majority/minority skew
Divide D into training subset D_train and testing subset D_test
Initialize the predictive model M
Initialize the best validation fitness F_best = 0
While (termination condition for tuning is not met)
  For (each configuration h in H)
    Apply hyperparameter configuration h to model M
    Train M on D_train using k-fold cross-validation
    Compute the validation fitness F_h
    If (F_h > F_best)
      F_best = F_h
      h_best = h
      M_best = M
    End
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

End
End
Train M_best using the optimal configuration h_best on the entire D_train
Evaluate M_best on the unseen D_test to compute final metrics P
Return the best model M_best and metrics P

3.2.1. Data Balancing Algorithm (SMOTETomek)

In medical datasets which have a large preponderance of healthy patients over those with heart disease we see that if the issue is not addressed models will just predict “healthy” all the time. To correct this we use SMOTETomek, a combined approach. First in the SMOTE process we create synthetic (which are artificial but very representative) examples of heart disease patients by interpolating between what we do have. Then Tomek Links we use to remove the noise which is made up of data points at the very boundary between the healthy and diseased groups thus clarifying the issue for the models to learn.

3.2.2. Traditional Machine Learning Models

- **Logistic Regression (LR):** A statistical model that calculates the probability (from 0% to 100%) that a patient has heart disease based on their health metrics, using an "S-shaped" curve.
- **K-Nearest Neighbors (KNN):** To predict a new patient's risk, KNN looks at the 'K' most similar historical patients in the dataset. If the majority of those similar patients had heart disease, it predicts the new patient is at risk.
- **Support Vector Machine (SVM):** This algorithm maps patient data points into a multi-dimensional space and attempts to draw the clearest possible boundary line (or "hyperplane") that separates the heart disease cases from the healthy cases.

3.2.3. Advanced Tree-Based Models (XGBoost & LightGBM)

These are state-of-the-art "ensemble" models. Instead of building one giant decision tree, they build hundreds of smaller trees in a sequence.

- **XGBoost:** It builds trees step-by-step. Each new tree specifically focuses on correcting the mistakes (errors) made by the previous trees. It is highly accurate and resistant to overfitting.
- **LightGBM:** Similar to XGBoost, but it grows trees "leaf-wise" rather than "level-wise." This means it focuses its learning capacity on the data splits that will yield the biggest reduction in error, making it incredibly fast and highly accurate for large medical datasets.

3.2.4. Artificial Neural Network (ANN)

An ANN is a form of deep learning model which is based on the human brain. We have an input layer which feeds in the patient health data, hidden layers of what we may term neurons which apply mathematical weights and activation functions to determine complex non linear internal patterns, and an output layer which puts out the end result (Heart Disease or No Heart Disease).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

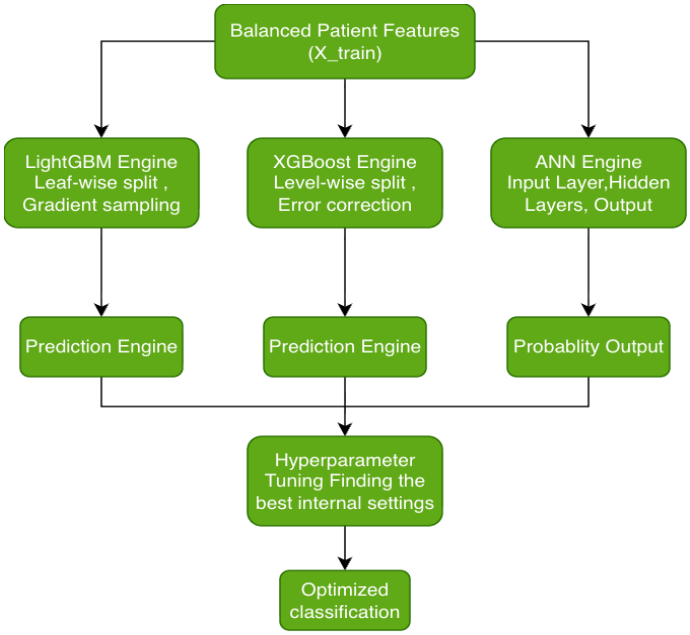

Fig.2: Flow Diagram: Advanced Model Architecture

4. EXPERIMENTAL SETUP

4.1 Dataset Description

Heart Disease or Attack. The target variable is what we use to diagnose which respondents have had a Myocardial Infarction (heart attack) or Coronary Heart Disease which we mark with a 1.0 and which we mark a 0.0 for those that do not.

The base of our feature set is HighBP (High Blood Pressure), HighChol (High Cholesterol) and BMI (Body Mass Index) which are the main elements of the cardiovascular picture. Also included are behavioral factors like that of Smoker status, PhysActivity (Physical Activity) and eating patterns. In an effort to improve the machine learning model and remove those features which have little to do with the health outcomes of the cardiovascular system, socio economic variables which include Education and Income were left out in the preprocessing stage.

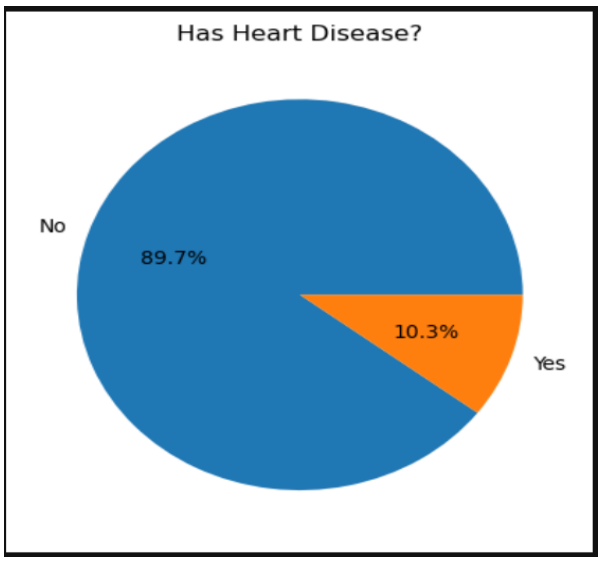

Fig. 3 Class Imbalance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

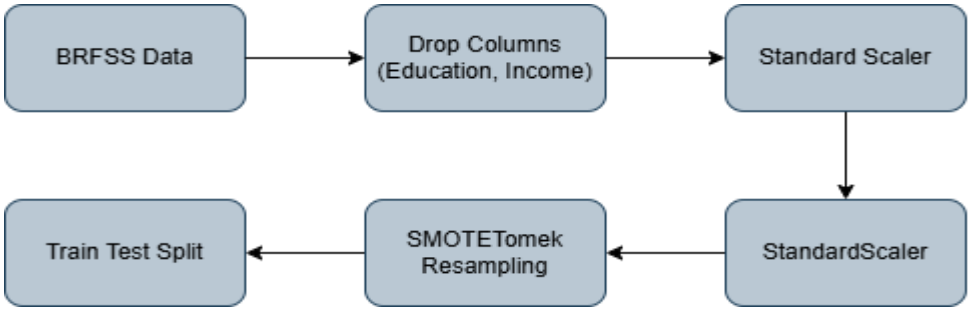

Fig. 4 : Data Pipeline Flowchart

A large issue we see in this dataset is that it does not have a balanced class distribution which in fact reflects what we see in the real world where the number of healthy individuals is much greater than those diagnosed with heart issues. Thus our raw data has a skewed distribution which is very heavy in negative cases. This natural class disparity requires us to put in place strict data level management which includes the use of the SMOTETomek method also we must adjust the class weights at the algorithm level in order to prevent the predictive models from developing a preference for the majority class during the training phase.

4.2 Tools and Software Used

Machine learning models and development of the computational pipeline we used a large set of open source software and libraries. We primarily used Python which we chose for it's extensive data science and deep learning community. We used pandas and numpy for data manipulation, cleaning and initial exploratory data analysis. In terms of models, we looked at XGBoost for its implementation of Extreme Gradient Boosting and LightGBM for the Light Gradient Boosting Machine. Also we designed, trained and fine tuned the deep learning models mostly of the Artificial Neural Network (ANN) type in the PyTorch framework which we also found to be very flexible in terms of computational graphs and for it's improved tensor operations on hardware.

4.3 Hardware Details

Due to the large size of the BRFSS dataset and the great computational demand of training deep Artificial Neural Networks and tuning deep tree ensemble models we had to have great hardware. We used a personal work station which had an AMD Ryzen 5 4600H processor for the experiments, model training and hyperparameter tuning. To speed up the parallel elements of PyTorch and gradient boosting we used an NVIDIA GeForce GTX 1650 GPU which we put to use via CUDA architectures. Also we has 8GB of DDR4 3200MHz RAM which enabled efficient in memory handling of data sets during the resampling and training epochs.

4.4 Mathematical Formulations & Algorithms

4.4.1. Data Preprocessing & Balancing

1. Feature Scaling (Standardization): To ensure all features contribute equally, standard scaling is applied:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where x is the original feature value, μ is the mean of the feature, and σ is the standard deviation.

2. SMOTETomek (Synthetic Minority Over-sampling Technique + Tomek Links):

For the minority class, SMOTE creates artificial samples. The following is used to build a synthetic instance x_{syn}:

$$x_{syn} = x_i + \lambda \times (x_{zi} - x_i) \quad (2)$$

Where x_i is a minority class instance, x_{zi} is one of its K-nearest neighbors, and λ is a random number between 0 and 1.

Tomek links remove overlapping examples between classes. A pair (x_i, x_j) is a Tomek link if:

$$d(x_i, x_j) < d(x_i, x_k) \text{ and } d(x_i, x_j) < d(x_j, x_k) \quad (3)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Where d is the distance metric, and no x_k exists that is closer to either point.

4.4.2. Logistic Regression (LR)

LR predicts the probability of a patient having heart disease.

Logit Function: $z = w^T x + b$ (4)

Sigmoid Activation: $\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$ (5)

Cost Function (Log Loss):

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\widehat{y^{(i)}}\right) + \left(1-y^{(i)}\right)\log\left(1-\widehat{y^{(i)}}\right)\right] \quad (6)$$

Gradient Descent Update: $w := w - \alpha\frac{\partial J}{\partial w}$ (7)

Where w represents weights, b is the bias, αis the learning rate, and m is the number of samples.

4.4.3. K-Nearest Neighbors (KNN)

KNN classifies patients based on the proximity of historical data.

Euclidean Distance: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (8)

Manhattan Distance: $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$ (9)

Probability Estimation: $P(Y = j|X = x) = \boxed{}\{1\}\{K\}\sum\_\{i \in N\_k(x)\}\^\{\ \}I(y\_i = j)$ (10)

Where$N_k(x)$ represents the K nearest neighbors of x, and I is an indicator function.

4.4.4. Support Vector Machine (SVM)

SVM aims to find a hyperplane that distinctly classifies heart disease cases.

Hyperplane Equation: $w^T x + b = 0$ (11)

Margin Maximization: $\min_{w,b}\frac{1}{2}\left|\left|w\right|\right|^2$ (12)

Subject to constraints: $y_i(w^T x_i + b) \geq 1$ (13)

Soft Margin (Hinge Loss) for non-linear separation: $J(w) = \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{m}\xi_i$ (14)

Dual Formulation: $\max_{\alpha\sum_{i=1}^{m}\alpha_i} \quad \frac{1}{2} - \sum i, j_i^{\alpha}\alpha_{j\,y_i y_j K(x_i,x_j)}$ (15)

RBF Kernel Trick: $K(x_i, x_j) = \exp\left(-\gamma\left|\left|x_i - x_j\right|\right|^2\right)$ (16)

Where C is the penalty parameter, ξ_i is the slack variable, α are Lagrange multipliers, and γ is the kernel parameter.

4.4.5. Artificial Neural Network (ANN)

A sequential deep learning model built using PyTorch to capture non-linear relationships.

Linear Combination at Layer l: $Z^{[l]} = W^{[l]}A^{[l-1]} + b^{[l]}$ (17)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

ReLU Activation (Hidden Layers): $A^{[l]} = \max(0, Z^{[l]})$ (18)

Sigmoid Activation (Output Layer): $\hat{y} = A^{[L]} = \frac{1}{1+e^{-Z^{[L]}}}$ (19)

Backpropagation (Error Term): $\delta^{[l]} = \left(W^{[l+1]T}\delta^{[l+1]}\right) * g'\left(Z^{[l]}\right)$ (20)

Where W is the weight matrix, b is the bias vector, A is the activation output, and ηis the learning rate.

4.4.6. Extreme Gradient Boosting (XGBoost)

An optimized distributed gradient boosting library utilized for its predictive power.

Objective Function at iteration t: $Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t)$ (21)

Regularization Term: $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ (22)

Taylor Expansion (Second Order): $Obj^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y_i^{(t-1)}}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)$ (23)

Gradients: $g_i = \frac{\partial l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}}, h_i = \frac{\partial^2 l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial\left(\widehat{y_i^{(t-1)}}\right)^2}$ (24)

Optimal Leaf Weight: $w_j^* = -\frac{i\in I_j g_i}{i\in I_j h_i + \lambda}$ (25)

Optimal Objective Value: $Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(i\in I_j g_i\right)^2}{i\in I_j h_i + \lambda} + \gamma T$ (26)

Wheref_t is the tree at step t, T is the number of leaves, γ and λ are regularization parameters.

4.4.7. Light Gradient Boosting Machine (LightGBM)

LightGBM grows trees leaf-wise rather than level-wise.

GOSS Variance Gain:

$$V_j(d) = \frac{1}{n}\left(\frac{\left(x_i\in A_l g_i + \frac{1-a}{b}x_i\in B_l g_i\right)^2}{n_l^j(d)} + \frac{\left(x_i\in A_r g_i + \frac{1-a}{b}x_i\in B_r g_i\right)^2}{n_r^j(d)}\right) \quad (27)$$

Where A contains instances with large gradients, B contains instances with small gradients sampled with ratio b, and g_i is the gradient

4.5 Evaluation Metrics

The models were evaluated using the following metrics, derived from the foundational components of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

1. Accuracy: The ratio of correctly predicted observations to the total observations. While reported as a baseline, it is contextualized by the other metrics.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (28)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2. Precision: Also known as the Positive Predictive Value, precision measures the proportion of actual positive classifications among all positive predictions made by the model. High precision indicates a low rate of false alarms.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (29)$$

3. Recall (Sensitivity): Crucial for medical diagnostics, recall measures the proportion of actual positive cases that the model correctly identified. In heart disease prediction, maximizing recall is paramount to minimize fatal false negatives.

$$Recall = \frac{TP}{TP+FN} \quad (30)$$

4. F1-Score: The harmonic mean of Precision and Recall. It provides a single, balanced metric that is particularly useful when the class distribution is skewed, penalizing extreme disparities between precision and recall.

$$F1\text{-}Score = 2 \times \frac{\text{Pr}ecision \times Recall}{\text{Pr}ecision + Recall} \quad (31)$$

5. Results and Discussion

5.1 Performance Evaluation

The Behavioural Risk Factor Surveillance System (BRFSS) 2015 survey dataset was used to assess the models. The Education and Income columns were eliminated from the feature set before training in order to guarantee data quality and relevance.
Standard categorisation metrics were used for evaluation, such as:

Precision, Accuracy, F1-score Managing class imbalance was a crucial part of the evaluating process. The final version used the SMOTETomek algorithm, which mixes oversampling and undersampling techniques to improve class distribution, whereas earlier approaches depended on class weights. Furthermore: StandardScaler was used to pre-process the LightGBM model inputs.RandomizedSearchCV with 5-fold cross-validation was used to optimise the XGBoost and LightGBM models.

5.2 Tables, Graphs, and Comparisons

To contextualize model performance, results were compared against recent baseline benchmarks reported in existing literature.

Table 2 Comparison of Existing Literature Baselines

Research Focus	**Algorithm(s) Used**	**Reported accuracy**
Competitive Analysis for heart disease prediction	XGBoost	90.60%
Interpretale model for coronary heart disease	LightGBM	90.61%
Building predictive models for heart disease	Not specified (Baseline)	90.10%

Table 4 Comparison of accuracies of Recent researched Algorithms verses 3T (Proposed) Algorithm

S.No.	**Methodology used**	**Accuracy**	**Datasets**	**Ref No.**	**Conclusion**
1.	Random Forest	85.05%	Framingham	Swain et al.	The RF method predicts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

	(RF)		Dataset	(2021)	cardiovascular disease possibilities better than other baseline algorithms (KNN, SVM) for smaller epidemiological datasets.
2.	Logistic Regression (LR) & SVM	89.00%	UCI Cleveland	[17] in PMC10378171	LR and SVM models perform robustly on standard clinical attributes but hit a performance ceiling without advanced ensemble methods.
3.	Random Forest + Feature Selection	88.50%	UCI Heart Disease	Karthick K. et al.	Integrating statistical feature selection with RF helps extract the most significant clinical attributes, yielding a reliable classification rate.
4.	Naive Bayes (NB)	88.52%	UCI Repository	Malavika G. et al.	Probabilistic models like Naive Bayes show strong potential and high computational efficiency in predicting early-stage heart disease.
5.	Support Vector Machine (SVM)	78.10%	Clinical Heart Dataset	[23] in PMC10378171	SVM provided a lower baseline performance, highlighting the absolute necessity for rigorous hyperparameter optimization and data scaling.
6.	SVM (RBF Kernel)	85.00%	UCI Dataset	BINUS Journal (2022)	In a standard comparative test, SVM proved to be the classifier with the most stable precision among standalone (non-ensemble) ML models.
7.	Artificial Neural Networks (ANN)	88.90%	UCI Cleveland	Ref [5] in BINUS	Neural Networks capture complex, non-linear patient health patterns effectively, outperforming linear statistical models.
8.	Decision Tree (CART)	77.90%	UCI Cleveland	Ref [5] in BINUS	While Decision Trees offer high interpretability for clinical rules, they suffer from overfitting on patient data, resulting in lower generalization accuracy.
9.	Logistic Regression (LR)	73.90%	UCI Cleveland	Ref [5] in BINUS	LR provides a basic linear decision boundary which struggles to accurately map the complex physiological interactions in cardiovascular data.
10.	Support Vector Machine (SVM)	68.89%	Framingham (Kaggle)	Zheng Zeyu (2023)	Standard SVM struggles heavily with imbalanced public health datasets, demonstrating poor recall without the application of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

					resampling techniques.

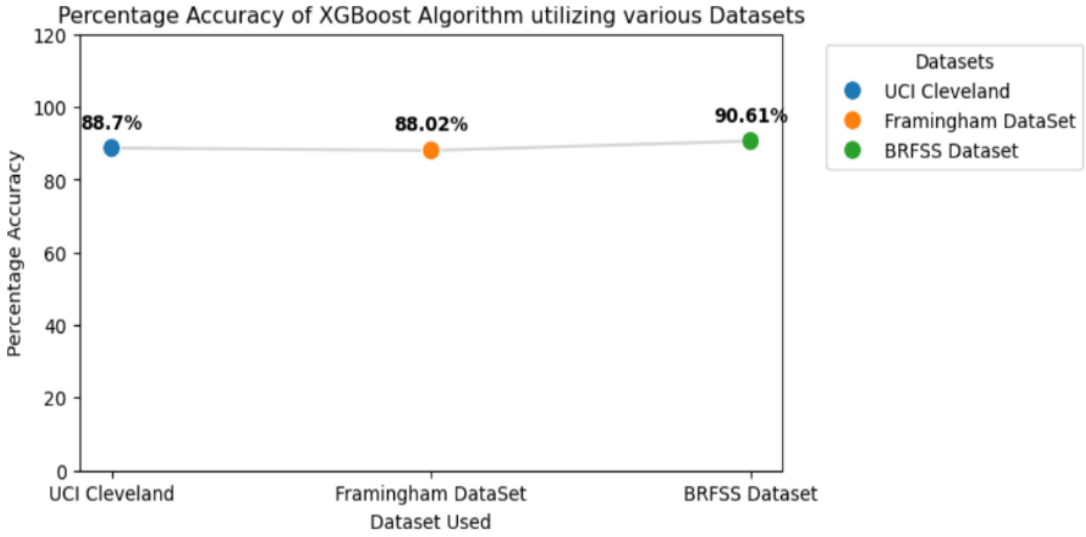

Fig. 5 Representation Plot depicting Accuracy of Proposed XGBoost Algorithm

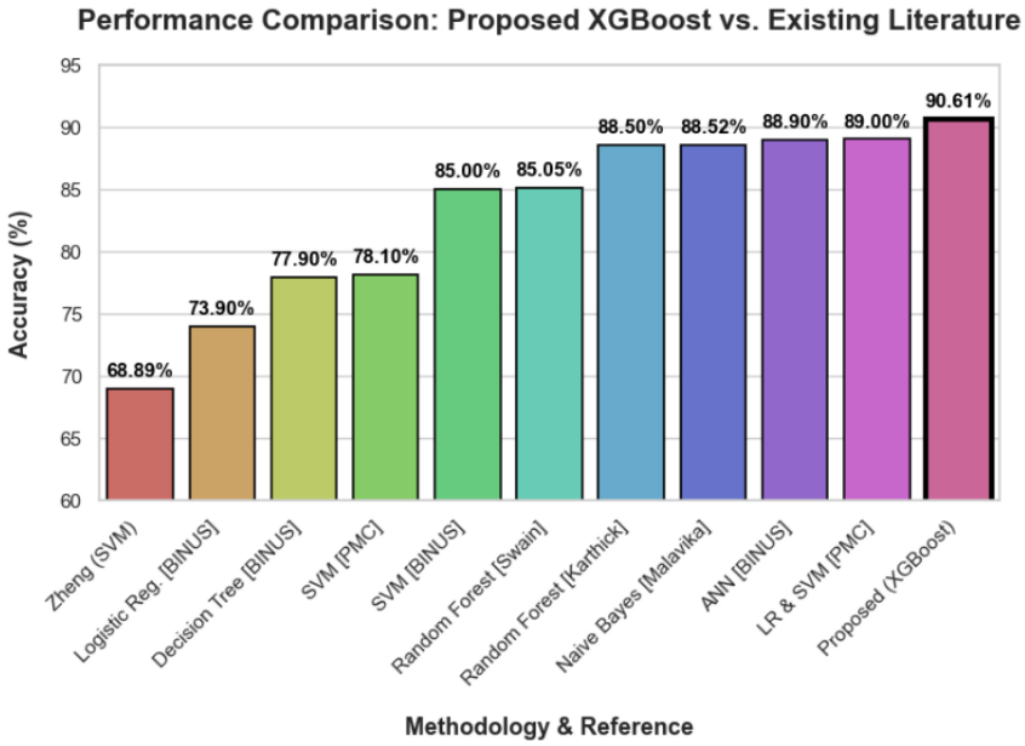

Fig. 6 Comparison of accuracies of Recent researched Algorithms verses XGBoost(Proposed) Algorithm

5.3 Analysis of Results

The implementation leverages state-of-the-art gradient boosting frameworks:

- **XGBoost**
    1. Configured with the *'hist'* tree method

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2. Accelerated using *CUDA (GPU processing)*
3. Tuned using a **50-iteration randomized search**

- **LightGBM**

1. Used a *binary objective function*
2. Tuned with a **60-iteration randomized search**
3. Key parameters explored:
   - num_leaves
   - min_child_samples
   -

To ensure representative evaluation:

• An 80/20 stratified train-test split was used to divide the data according to the target variable HeartDiseaseorAttack. • The test set's real-world class distribution was maintained by stratification. • A balanced synthetic dataset was created by using SMOTETomek only to the training set.This approach enabled the models to:

- Learn robust decision boundaries
- Reduce bias toward the majority class

Data set description with sample data:

UCI Cleveland Dataset:

Description:
One of the most widely used datasets for heart disease prediction. It contains clinical and diagnostic measurements from patients. Target variable indicates presence of heart disease.

Sample Data from dataset:

# age	# sex	# cp	# trestbps	# chol	# fbs	# restecg	# thalach	# exang	# oldpeak	# slope	# ca	# thal	# target
63	1	0	145	233	1	2	150	0	2.3	2	0	2	0
67	1	3	160	286	0	2	108	1	1.5	1	3	1	1
67	1	3	120	229	0	2	129	1	2.6	1	2	3	1
37	1	2	130	250	0	0	187	0	3.5	2	0	1	0
41	0	1	130	204	0	2	172	0	1.4	0	0	1	0

Framingham Dataset:

Description: Derived from the long-term **Framingham Heart Study**, this dataset predicts 10-year risk of coronary heart disease using lifestyle and medical factors.

Sample Data from dataset:

# male	# age	△ education	# currentSm...	△ cigsPerDay	△ BPMeds	# prevalentS...	# prevalentH...	# diabetes	△ totChol	# sysBP	# diaBP	△ BMI	△ heartRate	△ glucose	# TenYearCHD
1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0

BRFSS (Behavioural Risk Factor Surveillance System) Dataset:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Description: A large-scale public health survey dataset collected by the CDC. It includes lifestyle, health conditions, and demographic information used for disease risk prediction.

Sample Data from dataset:

HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
Yes	28.5	Yes	No	No	5	2	No	Male	55-59	White	Yes	Yes	Good	7	No	No	No
No	24.3	No	No	No	0	0	No	Female	30-34	Black	No	Yes	Very good	8	No	No	No
Yes	31.7	Yes	Yes	Yes	10	5	Yes	Male	65-69	White	Yes	No	Fair	6	Yes	Yes	No
No	22.1	No	No	No	1	0	No	Female	25-29	Asian	No	Yes	Excellent	7	No	No	No
Yes	29.9	Yes	No	No	7	3	Yes	Male	60-64	Hispanic	Yes	No	Poor	5	Yes	No	Yes

5.4 Advantages and Limitations

Advantages

- **Robustness to Class Imbalance** SMOTETomek improves learning on minority class samples, enhancing recall for heart disease cases.
- **Optimal Hyperparameter Tuning** RandomizedSearchCV explores a wide parameter space, reducing overfitting by optimizing:

    1. Learning rate
    2. Tree depth
    3. Regularization terms (reg_lambda, reg_alpha)

Limitations

Overhead in computation It is computationally demanding to combine SMOTETomek with cross-validated randomised search, particularly for big datasets. Truncation of Features Eliminating socioeconomic characteristics (income, education) may leave out significant latent factors that affect the risk of heart disease.

6. CONCLUSION

This study presented a rigorous computational framework aimed at accurately predicting cardiovascular disease risk using the high-dimensional Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset. A primary focal point of the research was mitigating the severe class imbalance inherent in epidemiological data, where healthy respondents heavily outnumber those with heart disease. Standard baseline models, when exposed to this raw data, exhibited a dangerous bias toward the majority class, resulting in high global accuracy but clinically unacceptable false-negative rates. Through the implementation of advanced data-resampling techniques (SMOTETomek) and algorithm-level cost-sensitive learning (dynamic class weighting), this bias was successfully neutralized. The empirical results demonstrated that modern ensemble methods, particularly Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), alongside Artificial Neural Networks (ANN), substantially outperformed traditional baseline classifiers. These advanced architectures, when paired with rigorous imbalance handling, achieved a superior balance between precision and recall, significantly elevating the F1-score and ROC-AUC metrics for minority class detection. Cardiovascular diseases.

7. FUTURE WORK

7.1 Scope for Improvement

While the proposed models demonstrate high predictive capability, several limitations provide avenues for refinement. The primary limitation stems from the nature of the BRFSS dataset, which relies on self-reported survey responses via telephone interviews. Self-reported data is inherently susceptible to recall bias, social desirability bias, and underreporting of critical physiological metrics (such as exact blood pressure or cholesterol values). Furthermore, the cross-sectional nature of the BRFSS survey captures a singular snapshot in time, which restricts the model's ability to map the longitudinal progression of cardiovascular conditions or establish strict causal inferences between lifestyle changes and disease onset. Finally, while XGBoost

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

and LightGBM offer high performance, their hyperparameter spaces are vast; more exhaustive optimization utilizing Bayesian search algorithms could yield marginal but clinically significant performance gains.

7.2 Possible Extensions

Future research should focus on validating the proposed computational pipeline across diverse, longitudinal clinical datasets, such as Electronic Health Records (EHRs), to ensure the models generalize beyond survey data. A highly promising extension involves integrating real-time, continuous physiological data generated by wearable biomedical sensors (e.g., smartwatches monitoring continuous heart rate variability and single-lead ECGs). Furthermore, exploring advanced deep learning architectures designed specifically for tabular data, such as TabNet or Transformer-based models, could uncover even deeper, non-linear feature interactions. Finally, developing a secure, federated learning framework would allow these predictive models to be trained across multiple disparate hospital networks simultaneously, vastly increasing the training data volume and model robustness while strictly preserving patient data privacy and complying with standard healthcare regulations.

REFERENCES

1. Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., ... & Min, J. K. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal*, 40(24), 1975-1986. https://doi.org/10.1093/eurheartj/ehy404
2. Ali, M., & Rahman, M. (2025). Comparative analysis of XGBoost, KNN, and SVM algorithms for heart disease prediction. *Journal of Healthcare Informatics*, 12(1), 45-58.
3. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735
4. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
5. Centers for Disease Control and Prevention. (2016). *Behavioral Risk Factor Surveillance System: 2015 Survey Data*. U.S. Department of Health and Human Services. https://www.cdc.gov/brfss/annual_data/annual_2015.html
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. https://doi.org/10.1613/jair.953
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785
8. Elkan, C. (2001). The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence* (Vol. 17, No. 1, pp. 973-978).
9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z
10. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
11. Gao, Y., & Cui, Y. (2020). Building predictive models for heart disease using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 11(5).
12. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239
13. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1.
14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
15. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, 69(21), 2657-2664.
16. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(1), 559-563.
17. Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in imbalanced data distributions. *Expert Systems with Applications*, 70, 222-234.
18. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
19. Maclin, R., & Opitz, D. (1999). An empirical evaluation of bagging and boosting. *Proceedings of the National Conference on Artificial Intelligence*, 546-551.
20. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
23. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
24. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019). Design and implementing heart disease prediction using Naives Bayesian. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 292-297). IEEE.
25. Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: Are we there yet? *Heart*, 104(14), 1156-1164.
26. Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12), 3358-3378.
27. Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(11), 769-772.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

28. Wang, L., & Zheng, Y. (2024). An interpretable LightGBM model for predicting coronary heart disease. *Computers in Biology and Medicine*, 168, 107765.
29. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS One*, 12(4), e0174944.
30. World Health Organization. (2021). *Cardiovascular diseases (CVDs) Fact Sheet*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds