INTERNATIONAL
HELLENIC
UNIVERSITY

Enhancing preventive healthcare: Identifying high-risk patients for cardiovascular diseases

Konstantina – Vasiliki Tompra

SID: 3308220023

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

JANUARY 2024

THESSALONIKI – GREECE

Enhancing preventive healthcare: Identifying high-risk patients for cardiovascular diseases

Konstantina-Vasiliki Tompra

SID: 3308220023

Supervisor:	Prof. Christos Tjortjis
Supervising Committee	Dr P. Koukaras
Members:	Dr L. Akritidis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

JANUARY 2024

THESSALONIKI – GREECE

Abstract

This dissertation was conducted as a part of the MSc in Data Science at the International Hellenic University.

The global fight against cardiovascular diseases (CVD) is experiencing a plateau in progress. One of the major causes of this issue, is that it is extremely difficult even for health practitioners to predict heart diseases as it is an intricate task, demanding a great amount of knowledge and experience. In such times, there exists a growing demand to integrate machine learning (ML) and data mining within the healthcare system, as by harnessing the wealth of available data, insights to society can be very beneficial.

This research successfully addresses a significant gap in the existing literature, by thoroughly examining both machine learning models and neural networks for CVD risk prediction based on personal lifestyle factors in a highly imbalanced real-life dataset. We trained multiple classifiers, including namely, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), CatBoost and Artificial Neural Networks (ANN). We used the Behavioral Risk Factor Surveillance System (BRFSS) 2021 Heart Disease Health Indicators dataset and to tackle the class imbalance challenge, we used methods such as Synthetic Minority Over Sampling Technique (SMOTE) Sampling, Adaptive Synthetic (ADASYN) Sampling, SMOTE-Tomek, and SMOTE-ENN.

Based on the findings, we conclude that hybrid models like SMOTE-ENN and SMOTE-Tomek outperformed the alternative sampling techniques in terms of the sensitivity metric. Our proposed implementation includes SMOTE-ENN coupled with CatBoost optimized through Optuna, achieving a remarkable 88% on recall and 82% on the AUC metric. Also, the ANN proposed, exhibited promising results, offering an additional layer of robustness in detecting positive cases of cardiovascular diseases.

Acknowledgements

At this point I would like to express my deep gratitude to my Supervisor, Professor Christos Tjortjis for his valuable guidance and support in every step of this project. His useful suggestions and his willingness to provide his time along with his constructive criticism are very much appreciated.

I would also like to extend my sincere gratitude to ph. D candidate Georgios Papageorgiou for his willingness to share his expertise and his insightful feedback anytime it was needed.

I want to dedicate this work to the memory of my beloved mother; her love and support will always remain to my heart.

Konstantina-Vasiliki Tompra

07/01/2024

Contents

ABSTRACT **3**

ACKNOWLEDGEMENTS **4**

CONTENTS **5**

LIST OF FIGURES **6**

LIST OF TABLES **7**

1 INTRODUCTION **9**

2 LITERATURE REVIEW **11**

2.1 THEORETICAL BACKGROUND 11

2.1.1 Cardiovascular Diseases (CVDs) *11*

2.1.2 CVD risk factors *14*

2.1.3 Diagnostic tests for CVD *15*

2.1.4 The role of data analytics and predictive modeling in preventive healthcare *16*

2.2 RELATED WORK 18

3 MATERIALS AND METHODS **23**

3.1 DATA COLLECTION 23

3.2 PROPOSED METHODOLOGY 24

3.3 EXPLORATORY DATA ANALYSIS 25

3.4 DATA PREPROCESSING 27

3.5 FEATURE ENGINEERING 28

3.6 EVALUATION METRICS 29

4 RESAMPLING TECHNIQUES **30**

4.1 SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE) 31

4.2 ADAPTIVE SYNTHETIC SAMPLING (ADASYN) 32

4.3 SMOTE COMBINED WITH EDITED NEAREST NEIGHBORS (SMOTE-ENN) 33

4.4 SMOTE COMBINED WITH TOMEK LINKS (SMOTE-TOMEK) 34

5 MACHINE LEARNING MODELS 35

5.1 LOGISTIC REGRESSION 35
5.2 DECISION TREES 36
5.3 RANDOM FOREST 36
5.4 GRADIENT BOOSTING 37
5.5 XGBOOST CLASSIFIER 37
5.6 CATBOOST CLASSIFIER 38
5.7 ARTIFICIAL NEURAL NETWORKS 38

6 EXPERIMENTAL RESULTS 39

6.1 MACHINE LEARNING IMPLEMENTATION 40
6.1.1 Results interpretation on raw data 40
6.1.2 Results interpretation with SMOTE 44
6.1.3 Results interpretation with ADASYN 47
6.1.4 Results interpretation with SMOTE-Tomek 49
6.1.5 Results interpretation with SMOTE-ENN 51
6.2 DEEP LEARNING IMPLEMENTATION 53

7 DISCUSSION 57

8 CONCLUSION AND FUTURE WORK 59

8.1 CONCLUSION 59
8.2 FUTURE WORK 60

BIBLIOGRAPHY 65

List of figures

Figure 1: Cardiovascular Diseases (CVDs). 11
Figure 2: The role of predictive modelling in preventive healthcare. 18
Figure 3: Flowchart of the proposed methodology. 25
Figure 4: Percentage of people having a heart disease. 26

Figure 5: Heatmap illustrating correlations among all features. 27
Figure 6: SMOTE working procedure. 31
Figure 7: ADASYN working procedure. 33
Figure 8: SMOTE - ENN working procedure. 34
Figure 9: Augmentation using SMOTE-Tomek. 35
Figure 10: Working process of Random Forest algorithm. 36
Figure 11: Working process of Gradient Boosting. 37
Figure 12: Progression of XGBoost from Decision Trees. 38
Figure 13: Typical Neural Network layout. 39
Figure 14: Confusion matrix of Logistic regression performance. 41
Figure 15: Trade-off between Accuracy and Recall. 43
Figure 16: Over sampling with SMOTE. 44
Figure 17: Confusion matrix of XGBoost&SMOTE. 47
Figure 18: Over sampling with ADASYN. 47
Figure 19: Hybrid resampling with SMOTE-Tomek. 49
Figure 20: Hybrid resampling with SMOTE-ENN. 51
Figure 21: Confusion matrix of the peak performance (CatBoost&SMOTE-ENN). 53
Figure 22: Architecture details of proposed ANN model. 55
Figure 23: Model Recall Performance for Optimal Combinations. 58

List of tables

Table 1: BRFSS Dataset description. 23
Table 2: Performance results on raw data. 41
Table 3: Performance results after optimization. 42
Table 4: Macro and Weighted average f1-score. 44
Table 5: Performance results after SMOTE. 45
Table 6: Performance results after optimization (SMOTE). 46
Table 7: Performance results after implementing ADASYN. 48
Table 8: Performance results after optimization (ADASYN). 49

Table 9: Performance results after implementing SMOTE-Tomek.....50
Table 10: Performance results after optimization (SMOTE-Tomek). 51
Table 11: Performance results after implementing SMOTE-ENN. 52
Table 12: Performance results after optimization (SMOTE-ENN). 53
Table 13: Performance results for the ANN. 56

1 Introduction

World Health Organization (WHO) reported that around 17.9 million people die each year due to cardiovascular diseases, making them a primary cause of death across the globe [1]. While CVD mortality rates have shown a decline over the past thirty years, this `positive trend has started to level off, and there is a potential risk of it reversing unless significant and coordinated actions are taken. It is a battle that calls for a transformative approach to preventive healthcare, where we must not merely react to illness but anticipate it, intercepting the threads of fate before they intertwine into a potentially tragic outcome.

The accurate prediction of CVD risk based on personal lifestyle factors plays a crucial role in enabling early intervention and implementing preventive measures. However, diagnosis is a major problem for practitioners as the nature of the CVDs is highly complex, and often confused with signs of aging. Thus, in the past few years, machine learning algorithms have emerged as valuable tools in this field, leveraging their capacity to uncover intricate patterns and interactions within datasets [47].

This research focuses on early and efficient detection of heart disease at higher accuracy levels using machine learning and deep learning algorithms, based on history of past patient records. More specifically, two over-sampling and two hybrid resampling algorithms (SMOTE, ADASYN, SMOTE-Tomek, SMOTE-ENN), along with six ML models (Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), XGBoost Classifier (XGB), CatBoost) and an Artificial Neural Network were used for our predictive approach.

The main contributions of our study are given below:

- Tackling the class imbalance issue inherent in real-world medical datasets by employing various resampling techniques to improve the performance of our models.
- Identifying the maximum compatibility of specific classification algorithms with corresponding statistical sampling methods.

- Improve the models' ability to identify positive cases (sensitivity) in such an imbalanced dataset, with CatBoost elevating from 4% of recall to an impressive 88%.

Ultimately, this study effectively addresses a noteworthy research gap by thoroughly exploring machine learning and deep learning models for CVD risk prediction based on personal lifestyle factors on a highly imbalanced dataset. By comparing model performance, resampling methods, identifying influential attributes, and investigating the impact of hyperparameter tuning, the study provides valuable insights for healthcare professionals and researchers. By shifting our focus from curative medicine to predictive medicine, we aspire to create a paradigm shift in healthcare, fostering a world where prevention is no longer an afterthought but an inherent component of our collective well-being.

2 Literature Review

The aim of this section is to provide a comprehensive review of the relevant corpora regarding cardiovascular diseases and the crucial role of preventive healthcare in treating them.

2.1 Theoretical Background

2.1.1 Cardiovascular Diseases (CVDs)

Cardiovascular diseases (CVDs) refer to a class of diseases that involve the heart or blood vessels [2]. They are a significant global health concern and a leading cause of death and disability worldwide. CVDs encompass various conditions that affect the heart and blood vessels, including coronary artery disease (CAD), heart failure, arrhythmias, valvular heart diseases, peripheral vascular disease, and deep vain thrombosis [2]. In figure 1, we can see illustrated the most common ones.

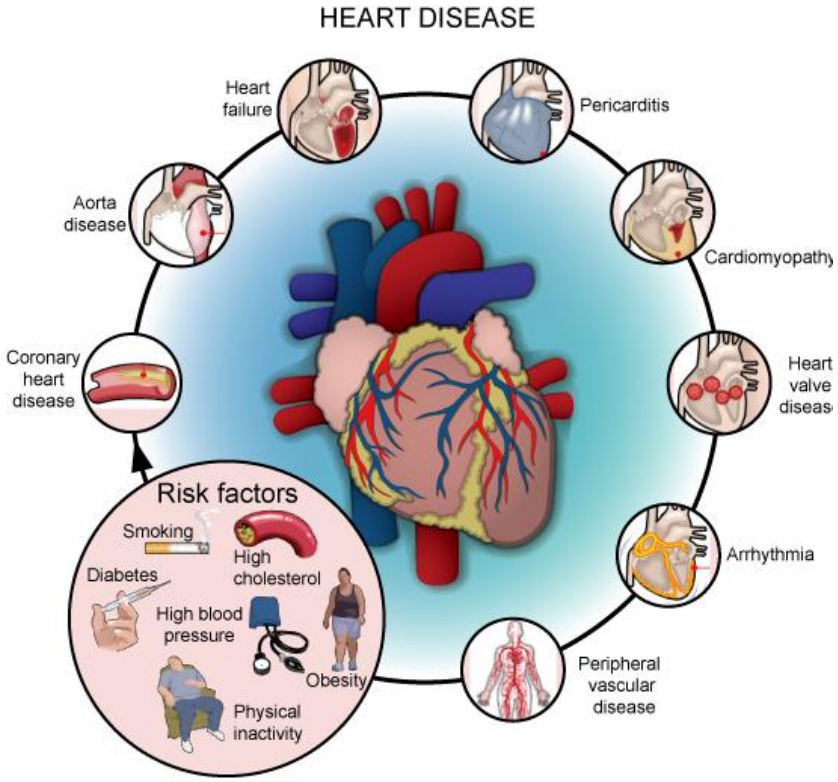

Figure 1: Cardiovascular Diseases (CVDs).

Coronary artery disease [3] is the most common type of cardiovascular disease and occurs when a buildup of plaque appears in the coronary arteries, which supply oxygen-rich blood to the heart muscle. Over time, the plaque can narrow or block these arteries, leading to decreased blood flow, angina (chest pains), and potential heart-related complications, such as heart attacks (myocardial infarctions). This situation is a medical emergency that requires urgent attention. However, you might have CAD for many years and not have any symptoms until you experience a heart attack. That's why CAD is considered a "silent killer". Risk factors for CAD include high blood pressure, high cholesterol levels, smoking, diabetes, obesity, and a sedentary lifestyle.

Heart failure [13],[14] occurs when the heart's ability to pump blood efficiently is compromised, leading to inadequate blood supply to meet the body's demands. It usually happens because the heart has become too weak or stiff and it needs some support to help it work better. Heart failure is a long-term condition that tends to get gradually worse over time and cannot usually be cured, but the symptoms can often be controlled for many years.

The primary signs of heart failure include:

- experiencing breathlessness even after light activity or while at rest
- constant fatigue accompanied by exhaustion during physical exertion
- feeling lightheaded or fainting
- swollen ankles and legs

Some people also experience other symptoms, such as a persistent cough, a fast heart rate and dizziness.

Symptoms can develop quickly (acute heart failure) or gradually over weeks or months (chronic heart failure).

Arrhythmias [15] refer to irregular heart rhythms that can either be too fast (tachycardia) or too slow (bradycardia). They occur due to disturbances in the heart's electrical system, which controls the heart's rhythm and rate. Normally, your heart beats in an organized, coordinated way. Issues with various parts of your heart — or even the blood your heart pumps — can affect your heart's normal rhythm. Having a normal heart rhythm matters because your heart supplies your whole body with nutrients and oxygen through the blood it pumps. Some types of arrhythmias are harmless and don't require treatment while others can put you at risk for cardiac arrest. Many are in between these two

extremes. A healthcare provider can tell you which type of arrhythmia you have and what kind of treatment you need, if any.

Heart valve disease [16] refers to any of several conditions that prevent one or more of the valves in your heart from working right. Left untreated, heart valve disease can cause your heart to work harder. This can reduce your quality of life and even become life-threatening. Conditions like aortic stenosis, mitral regurgitation, and mitral valve prolapse can impair the heart's ability to pump blood effectively. Valvular heart diseases can be congenital or acquired and may require surgical intervention in severe cases.

Peripheral Arterial Disease (PAD) [17] a common circulatory disorder in which narrowed arteries reduce blood flow to the limbs, typically the legs. It can result in symptoms like leg pain, numbness, or weakness, and in severe cases, it can lead to tissue damage or even amputation. It commonly occurs due to atherosclerosis, where plaque buildup narrows and blocks blood flow in the peripheral arteries. PAD can lead to pain, numbness, and non-healing wounds in the legs and feet. Left untreated, it may also increase the risk of heart attack and stroke.

Deep vein thrombosis (DVT, also called venous thrombosis) [18] occurs when a thrombus (blood clot) develops in veins deep in your body because your veins are injured or the blood flowing through them is too sluggish. The blood clots may partially or completely block blood flow through your vein. Most DVTs happen in your lower leg, thigh or pelvis, but they also can occur in other parts of your body including your arm, brain, intestines, liver or kidney.

Myocarditis [19], is an inflammation of the inner muscles of the heart caused by a variety of parasitic and microbial infections. It is a rare illness with only a few symptoms such as joint discomfort, limb swelling, or fever that cannot be diagnosed from the inside. Myocarditis is uncommon, but when it does occur, it is typically caused by an interior infection. Infections with microorganisms, fungi, parasites, viruses (most often, viruses that cause the flu virus, influenza, or COVID-19), or any other microorganisms can induce myocardial inflammation. Autoimmune diseases such as lupus, sarcoidosis, and others can trigger myocarditis due to the immune system’s ability to target any organ in the human body, along with the heart, and cause inflammation. Myocarditis can also be caused by drug usage, environmental exposure, or dangerous chemicals.

2.1.2 CVD risk factors

Cardiovascular disease (CVD) risk factors are conditions, behaviors, or characteristics that increase the likelihood of developing heart and blood vessel-related diseases. These risk factors can be modifiable, meaning they can be changed or managed, and non-modifiable, meaning they cannot be altered [2].

Modifiable Risk Factors

- High Blood Pressure (Hypertension): Elevated blood pressure puts extra strain on the heart and blood vessels, increasing the risk of CVD [44].
- High Cholesterol Levels: High levels of LDL cholesterol ("bad" cholesterol) and low levels of HDL cholesterol ("good" cholesterol) contribute to the buildup of plaque in the arteries, leading to atherosclerosis [44].
- Smoking [1]: Tobacco use damages blood vessels, reduces oxygen supply, and increases the formation of blood clots.
- Physical Inactivity [1]: A sedentary lifestyle is associated with obesity, high blood pressure, and other risk factors for CVD.
- Unhealthy Diet [1] : A diet high in saturated and trans fats, salt, and refined sugars contributes to the development of CVD.
- Obesity [1]: Excess body weight strains the heart and is linked to various risk factors like diabetes and high blood pressure.
- Diabetes [44]: People with diabetes have an increased risk of developing CVD due to the effects of high blood glucose levels on blood vessels.
- Stress: Chronic stress can affect behaviors and physiological processes, impacting heart health.
- Alcohol Consumption [1] : Excessive alcohol intake can raise blood pressure and contribute to heart muscle damage.

Non-modifiable Risk Factors:

- Age: As age increases, so does the risk of cardiovascular diseases.
- Gender: Men have a higher risk of CVD at a younger age, but women's risk increases after menopause.
- Family History: A family history of heart disease can increase an individual's risk.

- Ethnicity/Race: Certain ethnic groups have higher rates of specific cardiovascular conditions.

It is crucial to understand that the presence of one or multiple risk factors does not automatically mean one will develop cardiovascular disease [2]. Nonetheless, taking proactive measures such as adopting a healthier lifestyle, using medications as needed, and attending regular medical check-ups can substantially decrease the risk and enhance heart health. Seeking advice from healthcare experts can offer personalized assessments and guidance on preventive approaches.

2.1.3 Diagnostic tests for CVD

Diagnostic tests for cardiovascular diseases are essential for assessing the health of the heart and circulatory system. These tests help in the early detection, accurate diagnosis, and monitoring of various heart conditions. Here are presented the most common CVD diagnostic tests conducted by healthcare professionals.

Electrocardiogram (ECG or EKG) [2]: An ECG measures the electrical activity of the heart. It records the heart's electrical signals as waveforms, providing information about heart rhythm, rate, and any abnormalities, such as arrhythmias or signs of a previous heart attack.

Echocardiogram [4]: This is an ultrasound test that uses sound waves to create images of the heart's structure and function. It assesses the heart's chambers, valves, and pumping efficiency. Echocardiography helps in diagnosing conditions like heart valve problems, heart failure, and congenital heart defects.

Stress Test (Exercise ECG or Stress Echocardiogram) [4]: A stress test is performed while the patient exercises (e.g., walking or running on a treadmill or riding a stationary bike). It evaluates how the heart responds to physical stress and helps detect signs of reduced blood flow to the heart muscles, indicating possible coronary artery disease.

Cardiac CT Scan (Computed Tomography) [4]: Cardiac CT imaging uses X-rays to produce detailed cross-sectional images of the heart and blood vessels. It can assess coronary artery disease, heart structure, and function, as well as detect calcium deposits in the arteries (calcium scoring).

Cardiac MRI (Magnetic Resonance Imaging) [2][4]: Cardiac MRI uses powerful magnets and radio waves to create detailed images of the heart. It provides information about

heart function, tissue characterization, and can detect abnormalities in the heart's structure.

Coronary Angiogram (Cardiac Catheterization) [4]: This invasive procedure involves threading a catheter through the blood vessels to the coronary arteries. A contrast dye is injected, and X-rays are taken to visualize the coronary arteries and identify any blockages or narrowing (coronary artery disease).

Blood Tests [4]: Blood tests can measure specific biomarkers that indicate heart damage or strain. Common blood tests include cardiac enzymes (troponin, CK-MB) and B-type natriuretic peptide (BNP) or N-terminal pro b-type natriuretic peptide (NT-proBNP) for heart failure assessment.

Tilt table test [5]: Your provider will connect you to an ECG and blood pressure monitor. You will be strapped to a table that tilts you from a lying to standing position. This test is used to determine if you are likely to have sudden drops in blood pressure (orthostatic hypotension) while standing, or slow pulse rates with position changes. You might need this test if you often have fainting spells.

Holter Monitor [6]: It is a portable device worn by a patient that continuously records the heart's electrical activity (ECG) for 24 to 48 hours or longer. It helps diagnose irregular heart rhythms that may not be captured during a standard ECG.

Event Recorder [5]: Like a Holter monitor, an event recorder is a portable device used to record the heart's electrical activity, but it is typically used for a more extended period (up to several weeks or months). The patient activates the device when experiencing symptoms to capture any abnormalities.

Implantable loop recorder [5]:This device is about the size of a AAA battery. Your provider puts the device under the skin over the heart. The device monitors and records heartbeats for up to 3 years.

2.1.4 The role of data analytics and predictive modeling in preventive healthcare

As already explained, prevention is key to assist staying healthy and identifying potential health issues at an early stage before they lead to complications or become harder to manage. Unfortunately, uptake isn’t nearly as robust as it needs to be. One study from 2018 found that only 8% of adults in the United States who are 35 years and older received the preventive care recommended to them [7].

This disparity highlights the substantial gap between the potential benefits of preventive healthcare and its current utilization. In this context, the role of technology, data analytics, and predictive modeling in preventive healthcare has become increasingly vital. These strides in technology empower an approach to health management that is not only more vigilant but also highly personalized. Through the utilization of data analytics and predictive models, healthcare practitioners can meticulously examine extensive datasets encompassing an array of crucial information, spanning from an individual's medical history and lifestyle choices to genetic predispositions and beyond [8]. This comprehensive analysis enables the discernment of intricate patterns and risk factors, particularly those intricately tied to conditions such as cardiovascular diseases (CVD).

The prowess of these advanced tools goes beyond mere pattern recognition. It enables the precise identification of high-risk individuals before clinical symptoms manifest, thereby catalysing the deployment of targeted interventions and meticulously tailored treatment strategies [7]. Early detection holds the potential to bring about transformative changes in disease management and patient outcomes. By identifying the onset of health issues at their nascent stages, healthcare providers can initiate timely and personalized interventions that are tailored to the unique needs of an individual [9]. This not only enhances the effectiveness of treatments but also contributes to minimizing the potential complications, reducing the burden on healthcare resources, and ultimately improving the overall quality of life for patients [9]. Early detection, therefore, not only translates into medical benefits but also holds the promise of optimizing healthcare systems and promoting a healthier and more resilient population.

Medical imaging is also being extensively used in the analysis of X-rays, MRIs, CT scans, as machine learning and especially convolutional neural networks can accurately identify subtle abnormalities that may indicate the very early stages of many diseases like cancer, tumors, pneumonia etc.

Another outcome of particular significance is the potent synergy between technology and healthcare through wearable devices and remote monitoring technologies. These innovations empower real-time tracking of individuals' health metrics, facilitating timely interventions in response to any aberrations, and, notably, mitigating the necessity for frequent in-person appointments [10]. This has not only transformed the landscape of healthcare delivery but has also fostered an environment conducive to proactive health management. Another well-established and prevalent use of predictive analytics in healthcare is

identifying patients at high risk of hospital readmission [11]. Forecasting which patients may be readmitted after a hospital stay allows clinicians to adjust their post-hospitalization treatment strategies, noting that reducing readmissions saves money, preserves healthcare resources for new patients and improves patient outcomes [11].

Additionally, predictive analytics in the healthcare industry helps identify potential population health trends or outbreaks. The Lancet Public Health journal published a study that used predictive analytics to uncover health trends and found that unless alcohol consumption patterns will change in the US, alcohol-related liver diseases will rise, causing deaths [12]. When speaking of outbreak predictions, one can't help but ask, "could predictive analytics have foreseen the COVID-19 pandemic?". The answer is yes. BlueDot, a Canadian company building predictive analytics and AI solutions, issued a warning about the rise of unfamiliar pneumonia cases in Wuhan on December 30, 2019. Only nine days later, the World Health Organization released an official statement declaring the novel coronavirus emergence [12].

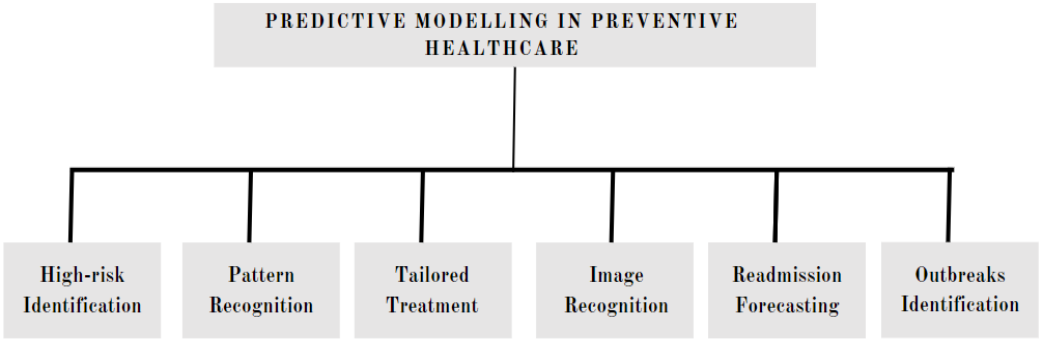

Figure 2: The role of predictive modelling in preventive healthcare.

However, as we embrace these technological advancements, it's important to address challenges related to data privacy, accuracy of predictive models, and equitable access to healthcare resources. Balancing the potential benefits of technology with ethical considerations will be instrumental in shaping the future of preventive healthcare.

2.2 Related work

Over the past few years, notable advancements have been made in the application of Data Mining and Machine Learning methods to predict cardiovascular diseases, with a

particular emphasis on early detection and prevention. This progress has been significantly influenced by pivotal studies in this field, and in this context, we will examine some of these studies, their approaches, results, and limitations.

Researchers have demonstrated that various machine learning or deep learning models have the potential to achieve high accuracy on predicting cardiovascular diseases. Weng et al. [26] from the very outset in 2017, evaluated four different models utilising clinical data sourced from over 300,000 homes in the United Kingdom. The outcomes revealed that among the methods examined, the neural network (NN) exhibited the highest accuracy in predicting cardiovascular disease, particularly when dealing with an extensive dataset under analysis, which highlights the need for more detailed and consistent electronic health data. Alqahtani et al. [27] devised an ensemble of machine learning (ML) and deep learning (DL) models achieving 88.70% accuracy for disease prediction, noting that in the end the ML Ensemble model was the most accurate. Gupta et al. [32], designed a machine intelligent framework (MIFH) for predicting heart diseases using the factor analysis of mixed data (FAMD) mechanism to derive features from the Cleveland dataset. In this study, not only an improved rate of sensitivity was achieved but also the MIFH system is able to return the best possible solution among all input predictive models considering performance criteria which can be very promising for the future.

Authors in [42], suggested a model that combines the Bagging ensemble learning method with decision tree and feature extraction with PCA and achieved an amazing 98.6% of accuracy on a realistic heart dataset. Paragliola and Coronato [33] formulated a predictive model tailored to anticipate the probability of cardiac events among hypertensive individuals, utilising ECG data as input. They innovatively combined a convolutional neural network with a long short-term memory network, resulting in a hybrid model. This integration harnessed time-series data to detect early increases in hypertension occurrences in individuals. Mohammed Nasir Uddin [34], focuses on developing an intelligent agent for predicting cardiovascular disease using an ensemble-based multilayer dynamic system. The proposed model employs five feature selection algorithms, along with an advanced ensemble learning model, and achieves high accuracy, with up to 94.16% accuracy and a 0.94 AUC value on a realistic heart dataset. What is worth noting is that this multilayer dynamic system can continue the classification process from one layer to another by enhancing its knowledge at each level to get the optimal result.

Al Ahdal et al. [35] in 2023 outperformed other machine learning algorithms mentioned in the literature section, achieving 96.7% of accuracy using the Random Forest classifier and 95.08% using the extreme gradient boost on the Cleveland dataset. Permatasari et al [53] achieved an 86% AUC score on predicting Diabetes Mellitus using the CatBoost classifier, and along with SHAP values they identified glycose levels and age as the most influential features. Most recently, Pasha and Mohamed [36], introduced an Advanced Hybrid Ensemble Gain Ratio Feature Selection (AHEG-FS) model that seeks to focus on improvement of the accuracy and AUC by selecting highly effective features while restoring relevant ones. Nine ML classifiers—AdaBoost, LR, classification via clustering (CVC), RF, k-nearest neighbour (KNN), support vector machine (SVM), boosted regression tree (BRT), naïve Bayes (NB), and stochastic gradient boosting (SGB)—are applied with the proposed AHEG-FS model, which is streamlined on medical datasets aimed at designing an innovative methodology for enhancing the prediction performance, and achieved an impressive 99% AUC after 46.15% features reduced. According to Ahmed et al. [40], methods like CatBoost, Random Forests, and Gradient Boosting can accurately foresee almost eight out of ten cardiac arrests. Asif et al. [20], also in 2023 achieved an impressive 98.15% accuracy on a Kaggle dataset revealing the power of ensemble methods like the Extra Tree Classifier on predicting heart diseases.

Sharma et al [37], suggested that deep neural networks should be further applied to address heart disease diagnosis, achieving 90% accuracy on the Cleveland dataset after also using Talos for the optimal hyper-parameters. Tick et al, [25] employs an ANN on the same dataset and evaluates its performance for different values of learning rate and number of neurons. The findings reveal that the highest accuracy of 80.6% is achieved with 0.25 learning rate and 25 neurons. In [38], a deep learning approach is suggested, along with the Isolation Forest algorithm for feature extraction and an improved 94.2% accuracy for the UCI dataset was achieved. Subramani Sivakannan et al. [39] later, developed a stacking model comprising both a base learner layer and a meta learner layer, yielding an impressive accuracy of nearly 96% on predicting the existence of a heart disease or not. These compelling outcomes underscores the potential of deep learning approaches in enhancing predictive performance.

However, the aforementioned studies on predicting heart diseases involve small and relatively balanced datasets. Here, we will address the problem of imbalance in an extensive dataset and attempt to identify which classification algorithms are the most

suitable for predicting heart diseases. Trigka et al. [28] innovatively employed stacking ensemble modelling by combining SVM, NB, and KNN with a 10-fold cross-validation synthetic minority oversampling technique (SMOTE) to tackle softly imbalanced datasets, resulting in a robust accuracy of 90.9%. Nishat et al. [29] employed the synthetic minority oversampling technique and edited nearest neighbour (SMOTE-ENN) data resampling technique, along with hyperparameter optimization and proved an evident enhancement of the classifiers performance especially on predicting the survival of patients with heart failure. Mahesh et al. [30], also utilised the Synthetic Minority Oversampling Technique (SMOTE) to cope with the problem of class imbalance as well as noise present in the Cleveland dataset and then with AdaBoost-Random Forest classifier achieved a 95.47% of accuracy in the early detection of heart disease. Dutta et al. [31] attempted to tackle the imbalance in the NHANES dataset with a two-step approach, involving the least absolute shrinkage and selection operator (LASSO) based feature weight assessment followed by majority-voting based identification of important features. Working on the 2015 BRFSS dataset, Teboul [45] tried to make some randomly selected splits of 60% not having heart disease to 40% having a heart disease and 50% not having heart disease to 50% of having a heart disease and highlighted Neural Networks, Gradient Boosting and AdaBoost as the most efficient models when it comes to the accuracy and AUC metrics. Authors in [46] decided to under-sample by random sampling the cases without CVD and aimed at reducing the consumption of medical resources and therefore the False Positive cases by building a 3-layered model that iteratively trains models and incorporate predictions from previous layers as features.

Lupague et al. [47] who utilized the 2021 BRFSS data that are used in our study, indicated that Logistic Regression should be more involved in the workflow for predicting cardiovascular diseases, as it correctly classified 79.18% of people with CVDs and 73.46% of people healthy and identified sex, diabetes, and general health of the patients as the most influential factors to predictions. In their research, Hairani et al. [51] achieved a remarkable 30.4% enhancement in model's sensitivity by integrating the SMOTE-Tomek algorithm with Random Forest. This outcome, as our study underscores, holds paramount significance in the context of heart disease prediction, as we believe is critical to minimize false negatives to ensure that high risk individuals will receive the medical attention they need, promptly.

3 Materials and methods

In this chapter, we introduce the problem we addressed, outlining our methodology. We provide a detailed description of the dataset used, the necessary data preparation, the feature engineering procedures, and the evaluation metrics used to assess the performance of our models.

3.1 Data Collection

The data collection process for this study involved accessing the 2021 annual Behavioral Risk Factor Surveillance System data (BRFSS) [43], a health-related telephone survey which was obtained from the Center for Disease Control (2021). The dataset, comprising 308,854 records with a total of 304 attributes, was accessed on a local machine for analysis and model development. However, not all these attributes were utilized to this specific study, as they were considered irrelevant. Therefore, a subset of 19 attributes was deliberately selected and was integrated into the construction of the predictive model, which aimed to identify high-risk individuals for cardiovascular diseases (CVD). The subset of the BRFSS dataset used, is displayed in Table 1, and it consists of 19 significant features.

Table 1: BRFSS Dataset description.

FEATURE	***DESCRIPTION***
General_Health	The general health condition of the respondent
Checkup	The period elapsed since the last time the respondent had a routine check-up
Exercise	Whether the respondent participated in any physical activities during the last month or not
Skin_Cancer	Whether the respondent had skin cancer or not
Other_Cancer	Whether the respondent had another kind of cancer or not
Depression	Whether the respondent had a depressive disorder or not
Diabetes	Whether the respondent had a diabetes or not
Arthritis	Whether the respondent had an arthritis or not

Sex	The respondent's gender
Age_Category	The category of age that the respondent fall into
Height_(cm)	The respondent's height measured in cm
Weight_(kg)	The respondent's weight measured in kg
BMI	The respondent's body mass index
Smoking_History	Whether the respondent had a smoking history or not
Alcohol_Consumption	The respondent's reported alcohol consumption
Fruit_Consumption	The respondent's reported fruit consumption
Green_Vegetables_ Consumption	The respondent's reported green vegetables consumption
Fried_Potato_ Consumption	The respondent's reported fried potatoes consumption
Heart_Disease	Whether the respondent reported a heart disease or not

3.2 Proposed methodology

Our study is aimed to predict cardiovascular diseases through the application of machine learning techniques. We employ a real-world dataset characterized by a significant class imbalance, highlighting the importance of maximizing the detection of false negative (FN) cases, intending to enhance preventive healthcare. Figure 3 provides a visual representation of our study's methodology.

Specifically, our methodology begins with preprocessing and feature engineering procedures, which include data cleaning, outlier detection, distribution checks, and data scaling. We also introduce additional features to unveil more intricate patterns within the data. Following data preparation, we address the class imbalance issue by employing two oversampling techniques (SMOTE and ADASYN) and two hybrid resampling methods (SMOTE-ENN and SMOTE-Tomek) on our training data.

Subsequently, we apply six machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and CatBoost) and construct an Artificial Neural Network (ANN). We then proceed to the optimization of the models through hyperparameter tuning.

The culmination of our study involves the presentation of the performance results obtained through the application of each resampling technique. Our primary focus lies on

identifying the most effective combination to maximize the sensitivity metric. This heightened emphasis on sensitivity is our contribution to the challenge of class imbalance within heart disease data, refining previous studies, which have often increased the false negative (FN) cases to achieve higher accuracy levels.

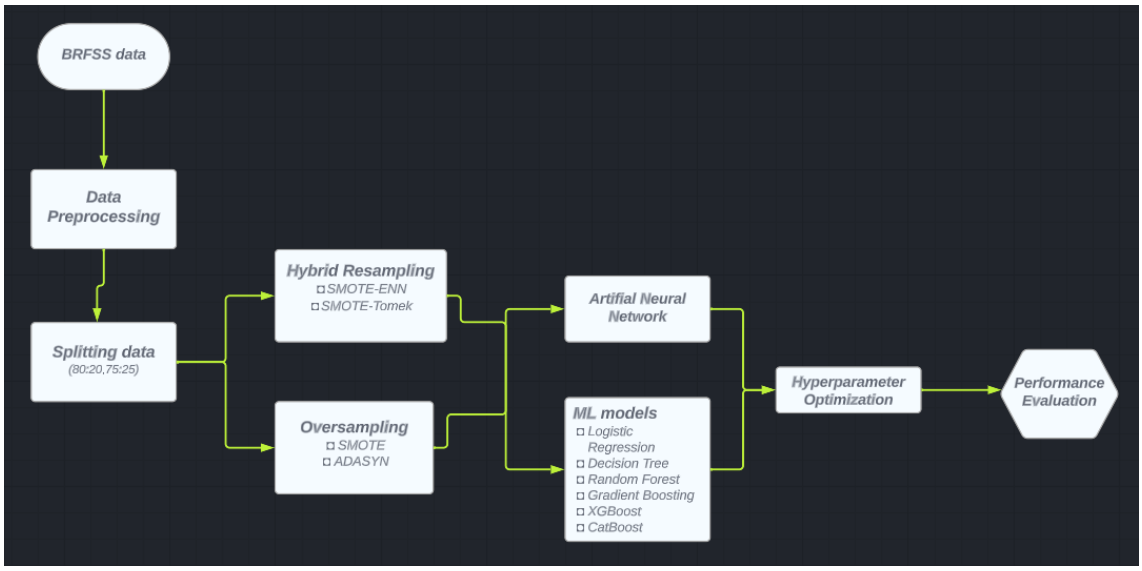

Figure 3: Flowchart of the proposed methodology.

3.3 Exploratory data analysis

Exploratory data analysis plays a vital role in predictive analytics, with the goal of providing insights into feature interactions, correlations, valuable patterns, and aiding in data understanding before making predictions. Various data analysis methods were employed to examine the BRFSS dataset and uncover insights about the relationships between different variables and the presence of heart disease. These methods include descriptive statistics, data visualization, and correlation analysis and some of these will be discussed below.

A preliminary statistical analysis was conducted so that we have a clear picture of the distribution of our data. Some important characteristics of our dataset include:

- There are slightly more females that males in the dataset.
- The dataset includes patients spanning from various age categories. Notably, the group aged 50-54 contains the highest number of patients, with the 55-59 and 60-64 categories following closely in terms of patient count. There is relatively less

participation from young individuals in the survey, which possibly implies that the predictive model may be more applicable to older demographics.

- The majority of patients assess their overall health as "Good," with "Very Good" being the next most frequently chosen option. Relatively fewer patients categorize their health as "Fair" or "Poor."
- Most patients underwent a checkup in the previous year.
- The majority of patients reported not suffering from Diabetes, Arthritis, Cancer or Depression.
- Most of the patients don't have a smoking history and do exercise regularly.

Next, it is very important to have a precise comprehension of the distribution of the target variable in the dataset. In the BRFSS data, we address a significant class imbalance, as shown in Figure 4, only 8.1% of the population participating in the survey reported having a heart disease, which can have serious adverse effects on our model. The model may be seriously biased towards the majority class of not having a CVD, leading to poor performance on identifying high-risk patients.

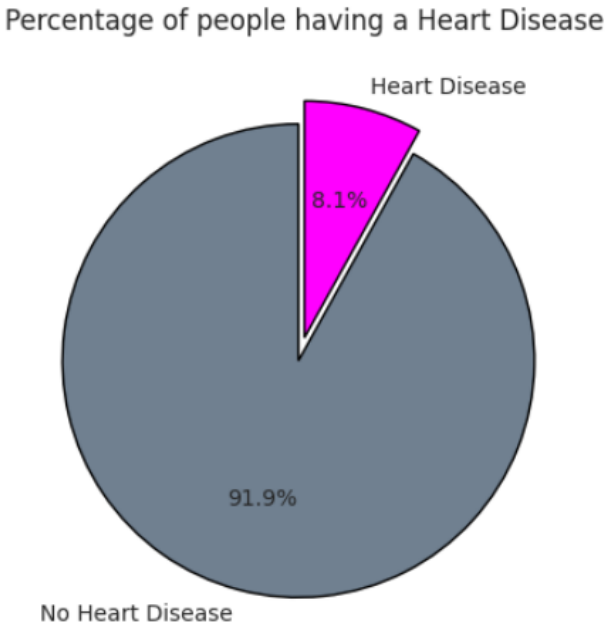

Figure 4: Percentage of people having a heart disease.

Next, we provide a heatmap that visually represents the correlations between all features using color encoding in a two dimensional format. It not only provides a clear visual representation of feature – target relationships, but also serves as a foundation for the detection of promising features for predictive models and offers insights into the factors

influencing CVD risk. Figure 4 shows how features are related to each other, with values closer to 1 indicating a robust positive relationship, values closer to -1 indicating a strong negative relationship, and values approximately 0 signifying the absence of a relationship. As we observe, General Health has a negative correlation with Heart Disease and Diabetes, which suggests that people who rated their general health as poor, are more likely to develop one or both of these diseases. Exercise also shows a negative correlation, suggesting that exercising can help reduce the risk of developing a disease. Age category appears to have a positive correlation with the target variable, which was anticipated, as it is non-modifiable risk factor, which has long been recognized as a critical determinant, with CVD incidence increasing as individuals grow older. Overall, we see that the most influential features for our prediction are General_Health, Age_Category, Diabetes, Arthritis and Exercise. This finding is also consistent with the established knowledge in the field of CVDs as research in this area has identified these factors as significant contributors to the risk and progression of a CVD.

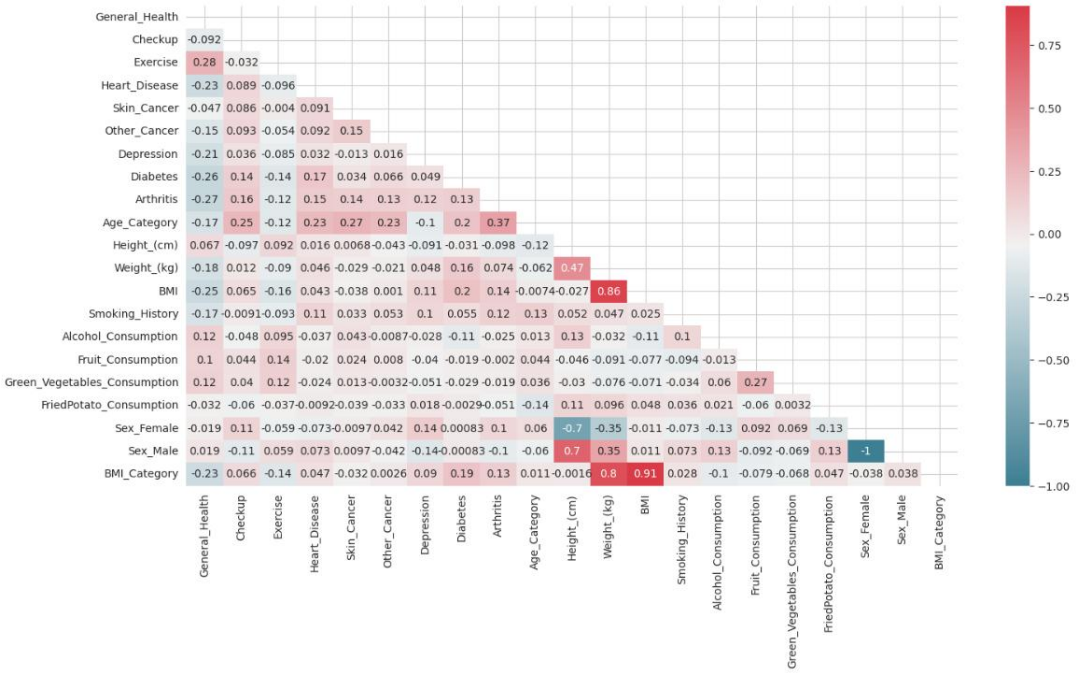

Figure 5: Heatmap illustrating correlations among all features.

3.4 Data preprocessing

The next step is the preprocessing of the data, which involves making the data more machine-readable and suitable for modeling.

The dataset doesn't contain any missing values and the output is binary, labeling with 0 people with no heart disease and with 1 people reporting having a heart disease. First, we had to remove 80 duplicated observations detected as they may introduce noise and inaccuracies in the dataset. When it comes to outliers, there were detected some quite high values in the Weight, Height, and BMI variables but we consider them as extreme values that are expected and potentially meaningful and therefore, we kept them in the dataset. Last, we normalized the input features with MinMaxScaler, so that all features are transformed in the [0,1] range, will all contribute equally to the model fitting, and avoid creating bias by using different scales.

3.5 Feature Engineering

We then proceed on some feature engineering to make data more informative and relevant to our predictive task. First, we employ binning on the "BMI" feature so that we can better interpret it and there may be some pattern identified. Literature shows that if your BMI is less than 18.5, it falls within the underweight range, from 18.5 to less than 25 seems to be in a healthy weight range, from 25 to less than 30, falls within the overweight range, and a BMI value higher than 30 indicates obesity [48]. Following, we created a variable "Overall_Diet", which provides a composite score of the individual's diet, taking into account the intake of green vegetables, fruits, and fried potatoes. The consumption of fruits and vegetables contributes positively to the score, whereas the consumption of fried potatoes detracts from it. Then, we aim at trying to identify a potential correlation between the bad habits of an individual and developing a heart disease, by creating the feature "Substance_Use", comprising the interaction with the combination of smoking and consuming alcohol. We use a different mapping for the smoking variable, labelling a smoker with -1 and a non-smoker with 0, so that higher negative values highlight a person making use of both tobacco and alcohol.

Next, we convert features like Heart_Disease, Skin_Cancer, Other_Cancer, Depression, Arthritis, Smoking_History, and Exercise that take only values Yes/No into their binary format. Then, we apply label encoding on the ordinal features like General_Health, BMI_Category, Age_Category so that we can preserve their ordinal nature, while we apply one-hot encoding on the rest categorical, nominal features as Sex and Diabetes so that

we prevent the model from making assumptions on the relationships between the categories.

3.6 Evaluation metrics

Accuracy

Accuracy represents the baseline performance of a classification model as it measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total ones [56]. It provides a general overview of the model's performance but can be misleading when used with imbalanced datasets as the one we are dealing with. A model may achieve high accuracy by correctly predicting the majority class while completely neglecting the minority class.

Recall / Sensitivity

Recall quantifies the model's ability to correctly identify positive instances out of all actual positive instances [23]. This metric holds paramount importance, especially in scenarios where the consequences of missing positive cases carry a substantial cost. In our specific case, where the primary objective is to identify and mitigate the risk of cardiovascular diseases, recall is our focus. As the repercussions of failing to detect individuals at risk are significant in this context, we prioritize the optimization of our model with a primary focus on maximizing recall. This emphasis ensures that our model excels in capturing a higher proportion of individuals with CVDs, aligning with the critical objectives of our healthcare application.

Precision / Specificity

Precision assesses the accuracy of positive predictions by calculating the ratio of correctly predicted positive instances to all predicted positive instances [30]. Particularly useful when minimizing false positives is critical.

F1-score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives [30]. It provides an assessment over the balance between precision and recall.

Confusion matrix

A table that summarizes the model's performance by categorizing instances into true positives, true negatives, false positives, and false negatives. It provides a detailed breakdown

of the model's errors and successes, facilitating a deeper understanding of its performance. The confusion matrix elements in our study are: true positive (TP), which were patients who had heart disease and were correctly diagnosed; true negative (TN), which were patients who did not have heart disease and were correctly diagnosed; false negative (FN), which were patients who had heart disease and were misdiagnosed; and false positive (FP), which were patients who did not have heart disease and were misdiagnosed [56].

Area Under the Curve (AUC)

AUC measures the area under the Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between true positive (TP) rate and false positive (FP) rate across various thresholds [57]. It offers a comprehensive evaluation of a model's ability to distinguish between classes.

4 Resampling Techniques

In the realm of data-driven decision-making, the quality and integrity of data are paramount. The success of predictive models, regardless of their application, relies heavily on the data they are trained on. One of the challenges encountered in real-world datasets is data imbalance, which presents a significant barrier to achieving accurate predictions and model generalization. Data imbalance occurs when the class distribution within the dataset is highly skewed, with the majority class overshadowing the minority class. Traditional machine learning models, when confronted with imbalanced data, tend to favor the majority class, and exhibit suboptimal performance, as they might overlook the subtle patterns within the minority class.

A commonly embraced strategy for addressing severely imbalanced datasets involves a technique known as resampling. This approach encompasses the removal of instances from the majority class, referred to as under-sampling, and/or the inclusion of additional instances from the minority class, which is known as over-sampling. The objective of this chapter is to provide a comprehensive understanding of the resampling techniques that have emerged as effective tools for mitigating the imbalance challenge. techniques include Synthetic Minority Over-sampling Technique (SMOTE), SMOTE combined with

Edited Nearest Neighbors (SMOTE-ENN), SMOTE combined with Tomek links (SMOTE-Tomek), and Adaptive Synthetic Sampling (ADASYN).

4.1 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is an oversampling technique that creates synthetic samples for the minority class. This approach mitigates the challenge of overfitting often posed by random over-sampling. It primarily operates within the feature space, constructing new instances through interpolation between closely located positive instances.

Working procedure

Initially, the total number of oversampled observations, denoted as N, is established. Typically, it is chosen to achieve a balanced binary class distribution of 1:1, although this can be adjusted on need [49]. The process commences by randomly selecting a positive class instance, followed by obtaining its K-nearest neighbours (typically set to 5 by default) [49]. Finally, N instances from this set of K neighbours are chosen to generate new synthetic instances. This is achieved by calculating the difference in distance between the feature vector and its neighbouring instances using a chosen distance metric [49]. Subsequently, this difference is multiplied by a random value in the range (0,1] and added to the original feature vector. This process is visually depicted in Figure 6.

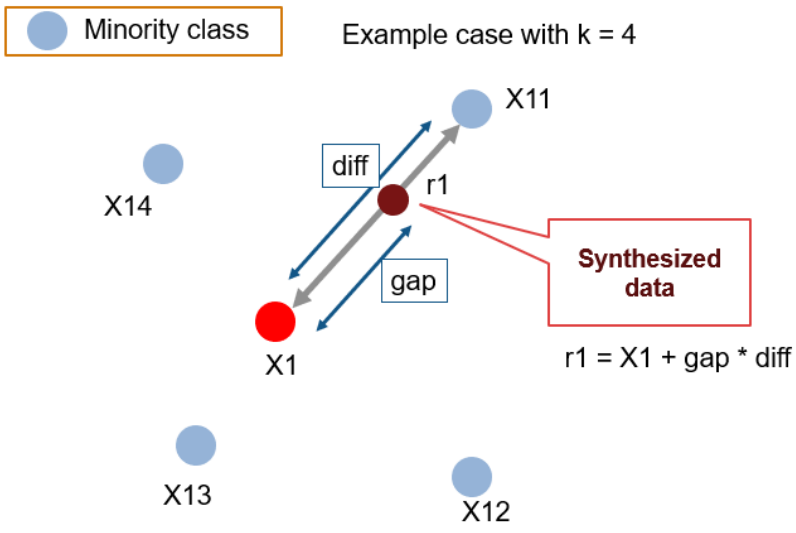

Figure 6: SMOTE working procedure.

4.2 Adaptive Synthetic Sampling (ADASYN)

ADASYN represents an extended version of the SMOTE algorithm. Like SMOTE, its primary goal is to boost the representation of the minority class by creating synthetic instances. However, the difference here is that it works in an adaptive manner, focusing on those instances that are more challenging to classify due to their proximity to the decision boundary. Instances that are harder to classify receive a higher oversampling rate, while those that are easier to classify receive a lower rate. Additionally, while SMOTE generates new data points strictly along straight lines between neighbouring points, the ADASYN algorithm delves deeper into the nearest neighbour area, by considering the majority class data points present within that region. Consequently, ADASYN generates synthetic samples only if there is a sufficient number of majority samples within the neighbouring region, ensuring a more context-aware oversampling technique [54]. This adaptability helps in maintaining a balance between boosting the minority class and preventing over-generalization.

Working procedure

From the dataset, we first determine the total number of instances in the majority class *(N-)* and the minority class *(N+)*. Then, we establish a predefined threshold value, d^{th} which serves as a limit for the maximum allowable class imbalance [49]. The total number of synthetic samples to be generated, is calculated as $G = (N^- - N^+)$ multiplied by β, where β is equal to (N^- / N^+) [49].

For every minority sample x_i, KNN's are obtained using Euclidean distance, and ratio r_i is calculated as Δ_i / k and further normalized as $r_x <= r_i / \sum r_i$ [49].

Thereafter, the total synthetic samples for each x_i will be, $g_i = r_x \, x \, G$. Now we iterate from 1 to g_i to generate samples the same way as we did in SMOTE [49].

Figure 7 represents the above procedure:

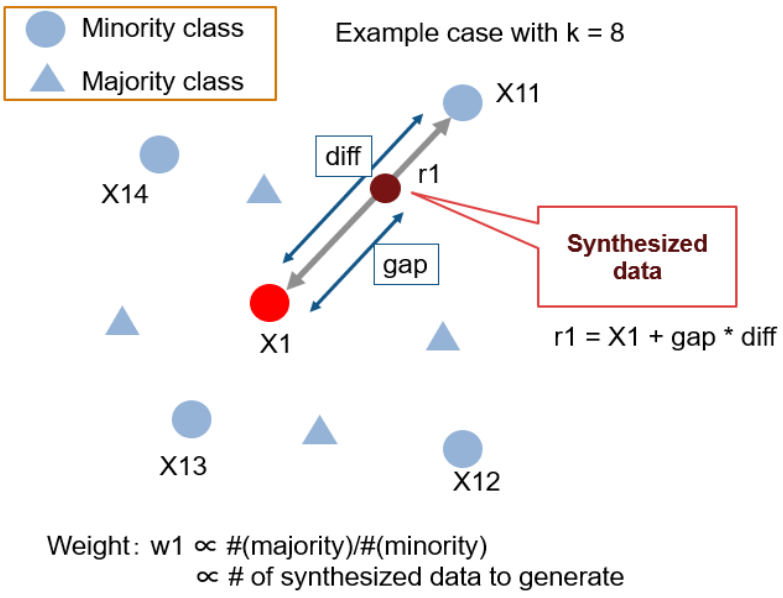

Figure 7: ADASYN working procedure.

4.3 SMOTE combined with Edited Nearest Neighbors (SMOTE-ENN)

Hybridization refers to the strategic combination of both under-sampling and over-sampling methods and has as primary goal the enhancement of the overall performance of the classifier models, specifically tailored to the datasets that have undergone these procedures.

SMOTE-ENN leverages the capabilities of both SMOTE and ENN to address class imbalance effectively. SMOTE augments the underrepresented class, ensuring a more balanced dataset, while ENN specializes in eliminating observations from both classes that deviate from their K-nearest neighbour majority class [54]. This hybrid approach ensures that the dataset is not only balanced but also free of noisy or misleading data points, thus significantly enhancing the quality and reliability of the dataset for subsequent model training and improved predictive performance.

Working procedure

The algorithm of ENN can be explained as below [50]:

Given the dataset with N observations, determine K, as the number of nearest neighbours. If not determined, then K=3.

Next, the algorithm identifies the K-nearest neighbours of a given observation within the dataset and determines the majority class among these neighbours.

If there is a disparity between the class of the observation and the majority class among its K-nearest neighbours, the algorithm proceeds to eliminate both the observation and its K-nearest neighbour from the dataset.

This process is repeated iteratively through steps 2 and 3 until the dataset attains the desired proportion of each class, achieving a balanced distribution.

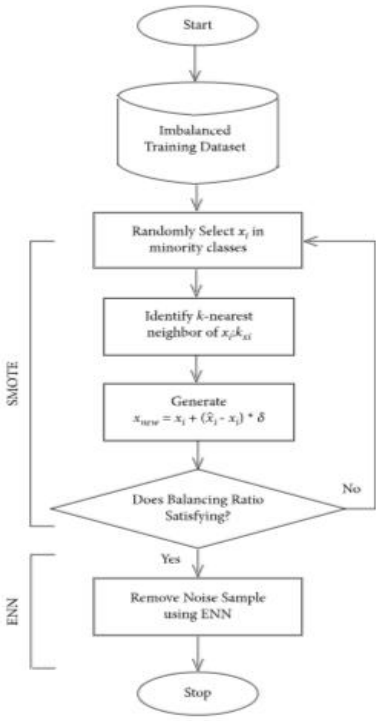

Figure 8: SMOTE - ENN working procedure.

4.4 SMOTE combined with Tomek links (SMOTE-Tomek)

SMOTE – Tomek is also a hybrid algorithm that works by first generating synthetic minority class instances to balance the data and then Tomek links, an under sampling technique, is responsible for pinpointing and removing noisy and borderline instances that reside near the decision boundary [51]. A Tomek link exists between two instances (datapoints) when they belong to different classes and are each other's nearest neighbors. In other words, they are a pair of instances from different classes that are very close to each other in the feature space. By removing these links, we eliminate noisy and ambiguous data points, leading to a cleaner and more balanced dataset.

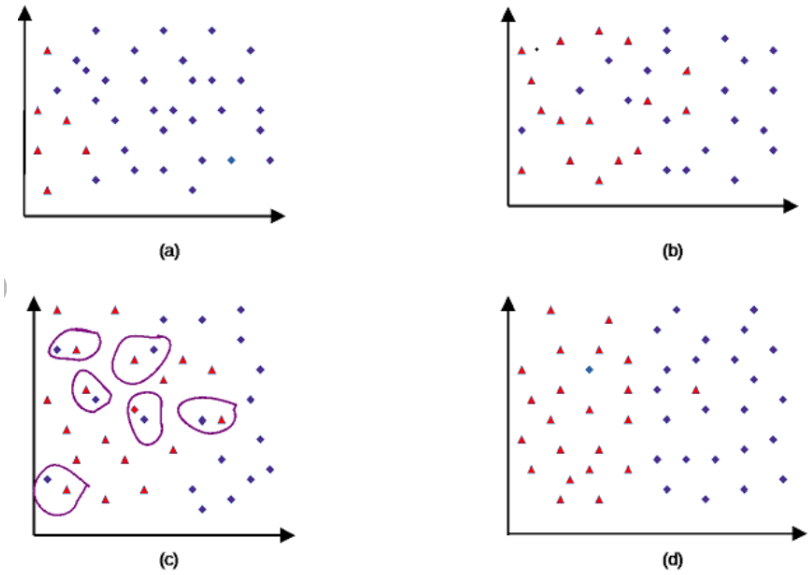

Figure 9: Augmentation using SMOTE-Tomek.

5 Machine Learning Models

Machine Learning, a branch of Artificial Intelligence, empowers applications that generate precise predictions, without explicit programming for predefined scenarios. In essence, Machine Learning epitomizes the capacity to glean insights and discern patterns autonomously, marking a transformative stride in the evolution of computational capabilities.

5.1 Logistic Regression

Logistic Regression is considered to be one of the most suitable models for predicting the likelihood of a target variable. This method employs the logistic or sigmoid function, characterized by an S-shaped curve, which transforms any real-valued input into a value within the range of 0 to 1[24]. To classify two classes, 0 and 1, a hypothesis $h(\theta) = \theta^T X$ is formulated, and the classifier's output is thresholded at 0.5 [23]. When the hypothesis value is above 0.5, it signifies a prediction of y = 1, indicating the presence of heart disease in the individual. Conversely, if the hypothesis value falls below 0.5, the prediction is y = 0, signifying a healthy person.

5.2 Decision Trees

Decision Tree is one of the most used supervised learning algorithms for both regression and classification tasks. These tree-like structures, with branches representing the decisions and the leaf nodes denoting the predicted outcomes offer a clear and interpretable way to make predictions on data [52]. The main objective of a decision tree is to encapsulate the training data in the most compact tree structure. They work by recursively partitioning the data into subsets based on the most informative features, ultimately creating a hierarchical set of rules to guide the decision-making process. Decision Trees can often be prone to overfitting when they become over-complex, however they can be enhanced through techniques like pruning and ensemble methods helping them to generalize better to unseen data.

5.3 Random Forest

The Random Forest algorithm is a powerful ensemble learning algorithm which builds upon the concept of decision trees by generating a multitude of individual decision trees during training. These trees are constructed with random subsets of the data and features, which introduces diversity and reduces the risk of overfitting and then the algorithm combines their predictions to make decisions [51]. Figure 10 illustrates its working process in classification.

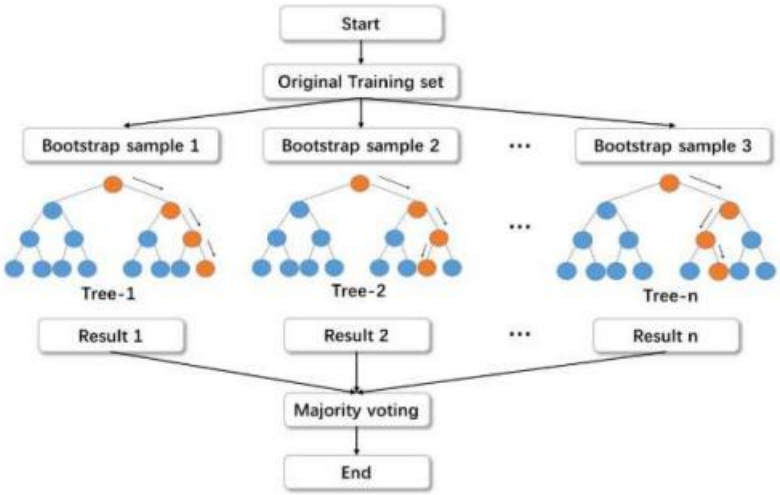

Figure 10: Working process of Random Forest algorithm.

5.4 Gradient Boosting

Gradient Boosting is a powerful ensemble learning technique, that operates by combining multiple weak learners, typically decision trees, into a strong predictive model. The key principle behind Gradient Boosting is to iteratively improve the model's performance by focusing on rectifying the errors introduced by the prior learners. The main benefit of gradient boosting lies in the continuous reduction of the residual error with each iteration [34].

Algorithm: Gradient Boosting

1. $F_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$
2. For m=1 to M do:
3. $\hat{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F_{x_i}}\right]_{F(x)=F_{m-1}(x)}, i = 1, \ldots, N$
4. $\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} [\hat{y}_i - \beta h(x_i; \alpha_m)]^2$
5. $\rho_m = \arg min \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; \alpha))$
6. $F_m(x) = F_{m-1}(x) + \rho h(x, \alpha_m)$
7. end For
8. end

Figure 11: Working process of Gradient Boosting.

5.5 XGBoost Classifier

Extreme Gradient Boosting is a scalable implementation of the Gradient Boosting framework employed for a wide range of tasks including both regression and classification. It focuses on optimization by employing various techniques like parallel processing and tree pruning to enhance speed and model accuracy [35]. It also incorporates regularization, to reduce overfitting, and allows users to define custom loss functions, offering a high degree of flexibility.

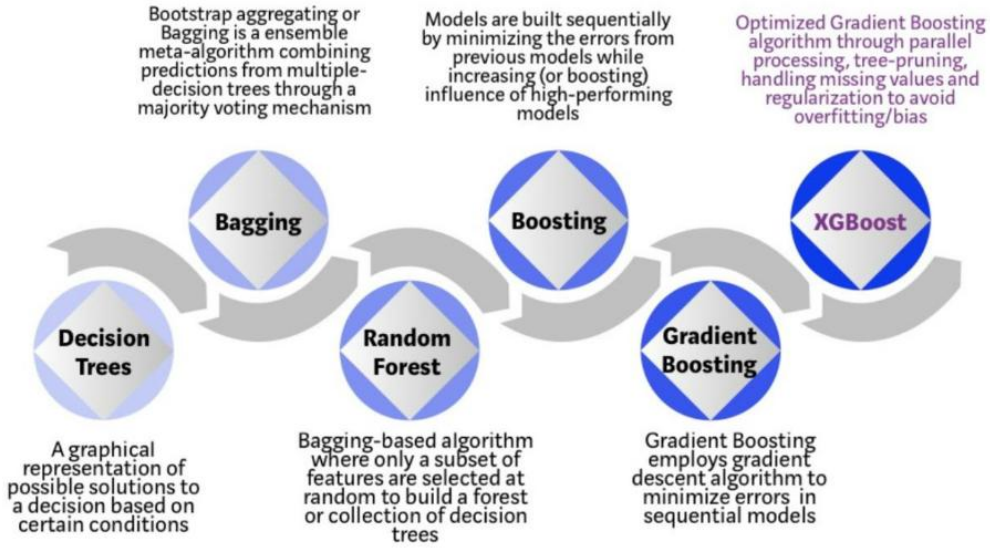

Figure 12: Progression of XGBoost from Decision Trees.

5.6 CatBoost Classifier

The Catboost algorithm is based on gradient boosted decision trees and usually outperforms other gradient boosting methods. Its unique approach involves the use of ordered target statistics and ordered boosting, making it particularly well-suited for handling categorical data in heterogeneous datasets [53]. This approach ensures that the model learns from categorical data without relying on one-hot encoding or label encoding. Each successive tree in the CatBoost algorithm is built with reduced loss compared to the previous ones [53]. It also reduces the need for extensive hyper-parameter tuning, uses categorical features directly and scalably, while it also allows specifying custom functions, features that make it a valuable tool for machine learning tasks.

5.7 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the human brain's intricate network of interconnected neurons. ANNs are comprised of interconnected layers of artificial neurons, or perceptrons, which serve as non-linear transformation units for input data, enabling them to carry out sophisticated tasks such as classification, regression, and pattern recognition. The defining characteristic of ANNs is their capacity for parameter adaptation through a process of iterative training, where the network adjusts its internal weightings to optimize its performance. Training the ANN

involves using a backpropagation network to adjust these weights, which occurs based on the disparity between predicted and actual outcomes. These weight updates are then propagated from the output (sink) to the input (source) layer in a feedforward network, aiming to minimize errors and produce output close to the target [21]. The fundamental component of an ANN is the artificial neuron, which computes its output by aggregating the inputs from the previous layer and applying an activation function, generating a numerical output within a predefined range, typically between 1 and -1, determined by the function's threshold [22]. In figure 13 below we can see a typical representation of feedforward neural network.

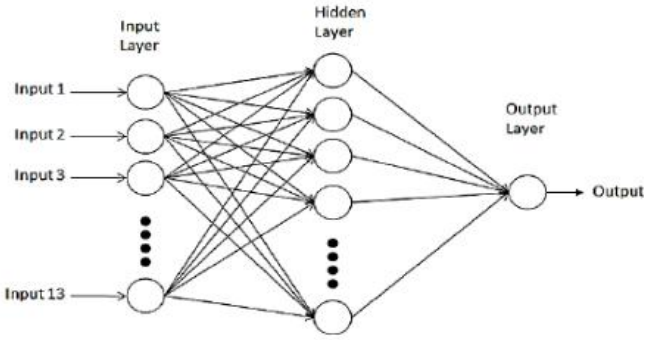

Figure 13: Typical Neural Network layout.

Although ANNs are usually associated with unstructured data like images, audio, and text, in this study we show that when working with structured data they can still provide valuable insights and make impressive predictions.

6 Experimental results

In this chapter we present the results of our research, including the performance results for each classifier after employing each one of the resampling techniques. Our split in all experiments was set to 70% for training each model and 30% for testing, stratified. Each of the resampling methods is employed only on the training data, so that we avoid data leakage from the test set; In this case we not only preserve the integrity of the test set, but we also provide more accurate evaluation of the model's generalization ability.

We evaluate the models' performance using metrics such as accuracy, recall, precision, F1-score, the area under the ROC curve (AUC) as well as the confusion matrix so that we have a detailed look on the classification of the observations.

6.1 Machine Learning implementation

6.1.1 Results interpretation on raw data

In the beginning of our research, we chose to employ six machine learning algorithms, namely, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB) and CatBoost on our raw data. The imbalance present poses a significant challenge for the models to correctly identify the positive observations, due to the limited representation of individuals who do have a heart disease in the dataset. Consequently, it biased the models to classify the majority of the observations as not having a heart disease, resulting in seemingly high accuracy while masking poor performance on the minority class. Addressing this challenge has been pivotal, and upon mitigation, the results exhibit notable improvements, as we will see in the next chapters.

We apply Stratified 5-fold Cross validation for all the models to obtain a more robust estimate of their performance, as by splitting the dataset into multiple folds and training/evaluating the model on each combination, we get a better sense of how well the models generalize to different subsets of the data.

The evaluation results of the trained ML models on the testing and the training set respectively, can be seen in Table 2.

An initial observation is that Decision Trees and Random Forest seem to have overfit to the training data, resulting into poor generalization to unseen data. We can also observe the high rate of accuracy (about 92%), almost all the models achieve while they seem to have a very poor performance on correctly predicting positive cases.

Table 2: Performance results on raw data.

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.92 / 0.93	0.03 / 0.06	0.47 / 0.51	0.06 / 0.12	0.80 / 0.84
Decision Tree	0.86 / 1.00	0.23 / 1.00	0.19 / 1.00	0.21 / 1.00	0.57 / 1.00
Random Forest	0.92 / 1.00	0.03 / 1.00	0.47 / 1.00	0.06 / 1.00	0.80 / 1.00
Gradient Boosting	0.92 / 0.92	0.05 / 0.05	0.49 / 0.55	0.09 / 0.09	0.83 / 0.84
XGBoost	0.92 / 0.92	0.05 / 0.09	0.46 / 0.75	0.10 / 0.17	0.83 / 0.88
CatBoost	0.92 / 0.93	0.04 / 0.11	0.47 / 0.84	0.09 / 0.19	0.83 / 0.87

This low rate of recall, although accompanied by a high accuracy rate, raises concerns about the models overlooking a significant portion of positive cases—a matter of paramount importance in predictive healthcare analytics. For example, as seen in Figure 14, the confusion matrix for Logistic Regression that performed a bit better, reveals 7247 missed positive cases. In a real-life scenario, such oversights could have profound consequences. Therefore, we've decided to prioritize increasing the sensitivity of our models.

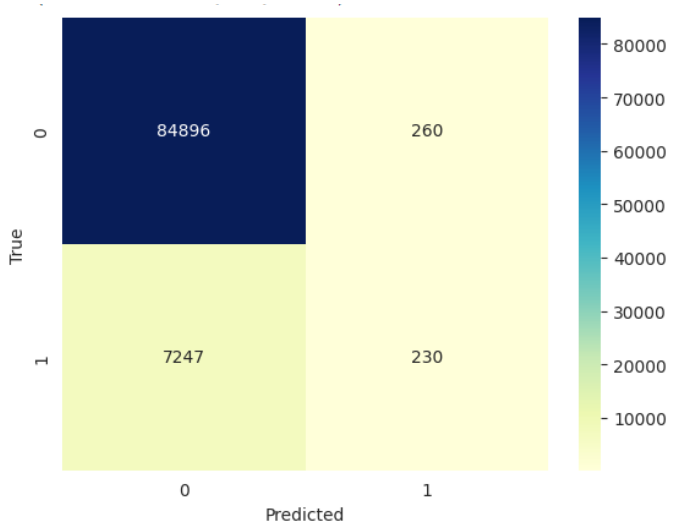

Figure 14: Confusion matrix of Logistic regression performance.

While default values often yield satisfactory results, the art of hyperparameter tuning unveils the potential for more accurate predictions. By reviewing the documentation of each algorithm, the bibliography, and by using some optimizing algorithms, we tried to find the right parameter grid to improve our models' performance.

We conducted an exhaustive grid search utilizing the GridSearchCV algorithm to optimize hyperparameters for Logistic Regression, Decision Trees, and Random Forest. While effective, this approach proved time-consuming. Following a meticulous exploration, we determined that the hyperparameter set [C = 0.01, penalty = 'l2', solver = 'liblinear'] demonstrated optimal performance for Logistic Regression. For Decision Trees, the parameters [criterion = 'entropy', max_depth = 10, min_samples_leaf = 2, min_samples_split = 10] yielded the best results, while Random Forest showcased peak performance with [n_estimators= 475, min_samples_split= 5, min_samples_leaf= 1, max_depth= 70, bootstrap= False].

Transitioning to boosting algorithms, we employed Optuna [55] to efficiently select the most effective hyperparameters. Notably, Gradient Boosting exhibited optimal performance with the set [n_estimators = 115, max_depth = 6, min_samples_split = 14, min_samples_leaf = 7], while CatBoost demonstrated superiority with [iterations=982, learning_rate = 0.014663724595555972, depth=9, l2_leaf_reg= 7.765080164412087, auto_class_weights= 'Balanced'].

Intriguingly, after multiple trials, XGBoost surpassed the results obtained through grid searches when assigned the hyperparameter set [scale_pos_weight = sum(y_train==0) / sum(y_train==1), eval_metric='logloss', use_label_encoder=False] which was manually found. This nuanced exploration underscores the importance of adaptive approaches in uncovering optimal configurations for machine learning models.

The performance results that the models achieved can be seen in Table 3.

Table 3: Performance results after optimization.

Model	***Accuracy/ Training Accuracy***	***Recall/ Training Recall***	***Precision/ Training Precision***	***F1 – score/ Training F1-score***	***AUC/ Training AUC***
Logistic Regression	0.92 / 0.92	0.05 / 0.05	0.52 / 0.53	0.08 / 0.09	0.83 / 0.84
Decision Tree	0.92 / 0.92	0.05 / 0.07	0.44 / 0.61	0.09 / 0.13	0.80 / 0.86
Random Forest	0.92 / 1.00	0.04 / 1.00	0.43 / 1.00	0.08 / 1.00	0.81 / 1.00
Gradient Boosting	0.92 / 0.92	0.04 / 0.07	0.49 / 0.71	0.08 / 0.12	0.83 / 0.86
XGBoost	0.74 / 0.76	0.75 / 0.87	0.20 / 0.23	0.32 / 0.37	0.82 / 0.89
CatBoost	0.74 / 0.78	0.77 / 0.85	0.20 / 0.22	0.33 / 0.35	0.83 / 0.87

Upon examination of the results, it becomes evident that there is negligible variance among the outcomes for the remaining models, except for XGBoost and CatBoost. Notably, these two models exhibit a noteworthy increase in the achieved recall, with CatBoost reaching an impressive 77%, while preserving an excellent accuracy rate. This substantial escalation from the initial 4% underscores the significant impact of tuning these models with the right hyperparameters. It demonstrates that, when finely tuned, CatBoost can identify a considerable number of individuals prone to cardiovascular diseases (CVD) with a level of accuracy that goes beyond mere satisfaction. The dual-axis visualization in Figure 15 offers insights into how each model navigates the trade-off between overall accuracy and the adeptness in capturing positive cases, recall.

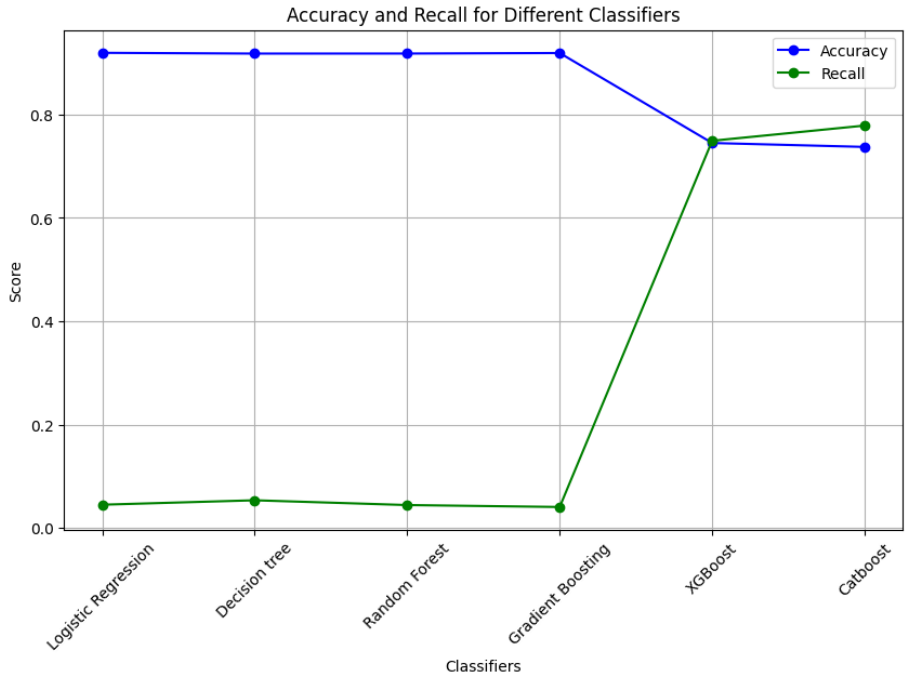

Figure 15: Trade-off between Accuracy and Recall.

Analysing also both the weighted average and the macro average F1-score as seen in Table 4, provides valuable insights into the model's performance with and without accounting for the proportion of each class. Notably, the macro average F1 score, even in the best-case scenario, is only 0.58. This suggests that the model's performance is not as strong when considering both classes equally, regardless of their imbalance.

Table 4: Macro and Weighted average f1-score.

Model	*Macro avg F1-score*	*Weighted avg F1-score*
Logistic Regression	0.52	0.89
Decision Tree	0.53	0.89
Random Forest	0.52	0.89
Gradient Boosting	0.52	0.89
XGBoost	0.58	0.80
CatBoost	0.58	0.80

6.1.2 Results interpretation with SMOTE

Next, in our journey to tackle the imbalance in the dataset, we employ the SMOTE algorithm, which introduces synthetic instances into the minority class in order to increase its representation in the data.

As we may see in Figure 16, SMOTE generates a sufficient number of observations to ensure an equal count between the two classes, resulting in a dataset comprising of 397.294 observations.

```
from imblearn.over_sampling import SMOTE

print("Before OverSampling- counts of label '1': {}".format(sum(y_train==1)))
print("Before OverSampling- counts of label '0': {} \n".format(sum(y_train==0)))

sm = SMOTE(random_state=42, k_neighbors=5)
X_resampled, y_resampled = sm.fit_resample(X_train,y_train)

print("After OverSampling with SMOTE - '1': {}".format(sum(y_resampled==1)))
print("After OverSampling with SMOTE - '0': {}".format(sum(y_resampled==0)))
```

```
Before OverSampling- counts of label '1': 17494
Before OverSampling- counts of label '0': 198647

After OverSampling with SMOTE - '1': 198647
After OverSampling with SMOTE - '0': 198647
```

Figure 16: Over sampling with SMOTE.

The performance results of the trained models on the test set in contrast with their performance on the training set after applying SMOTE in the dataset, can be seen in Table 5.

Table 5: Performance results after SMOTE.

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.70 / 0.70	0.64 / 0.65	0.16 / 0.17	0.26 / 0.26	0.74 / 0.75
Decision Tree	0.84 / 0.99	0.27 / 0.99	0.17 / 1.00	0.21 / 1.00	0.58 / 1.00
Random Forest	0.70 / 0.70	0.64 / 0.65	0.16 / 0.17	0.26 / 0.26	0.78 / 1.00
Gradient Boosting	0.83 / 0.84	0.39 / 0.40	0.21 / 0.22	0.27 / 0.28	0.77 / 0.77
XGBoost	0.91 / 0.91	0.11 / 0.16	0.31 / 0.42	0.17 / 0.23	0.80 / 0.84
CatBoost	0.91 / 0.92	0.08 / 0.14	0.40 / 0.62	0.14 / 0.23	0.81 / 0.87

The application of SMOTE demonstrates a remarkable enhancement in the models' capacity to identify individuals at risk of heart disease. Specifically, the recall metric exhibits a substantial surge, escalating from 3% to an impressive 64% when employing Logistic Regression or Random Forest algorithms. This heightened sensitivity suggests an improvement in correctly identifying positive cases, a critical aspect in the context of heart disease prediction.

However, it is essential to note the trade-offs accompanying this improvement. While the Recall metric experiences a notable boost, accuracy and precision witness a decline. This suggests that while the models become more adept at capturing instances of heart disease, there is a corresponding increase in false positives and a potential reduction in overall predictive accuracy.

Interestingly, all three Boosting algorithms don't exhibit a parallel enhancement in performance with the introduction of SMOTE. Despite the synthetic data augmentation, they appear to maintain the level of performance they did before over sampling the dataset.

These nuances in performance across different algorithms underscore the complexity of utilising SMOTE and its impact on various metrics.

We then proceed to fine tune our models, and the results achieved are presented in Table 6.

Table 6: Performance results after optimization (SMOTE).

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.70 / 0.70	0.64 / 0.65	0.16 / 0.17	0.26 / 0.26	0.74 / 0.75
Decision Tree	0.84 / 0.97	0.27 / 0.78	0.17 / 0.94	0.20 / 0.86	0.58 / 1.00
Random Forest	0.70 / 0.97	0.64 / 0.65	0.16 / 0.17	0.26 / 0.26	0.78 / 1.00
Gradient Boosting	0.83 / 0.84	0.39 / 0.40	0.21 / 0.22	0.27 / 0.28	0.77 / 0.78
XGBoost	0.67 / 0.68	0.80 / 0.89	0.17 / 0.19	0.28 / 0.31	0.80 / 0.85
CatBoost	0.70 / 0.71	0.79 / 0.80	0.18 / 0.20	0.30 / 0.33	0.81 / 0.87

The results presented in Table 6 indicate that, despite a meticulous grid search to optimize their hyperparameters, there was no improvement in the performance of Logistic Regression, Decision Trees, and Random Forest, and on top of that the last two are likely to have overfit the training data.

Nevertheless, our boosting algorithms have surpassed our expectations. CatBoost achieves an impressive 79%, and XGBoost attains a noteworthy 80% recall. Moreover, both models maintain a robust 80% on the AUC metric, emphasizing their resilience and discriminative prowess. As illustrated in the confusion matrix presented in Figure 17, the optimized XGBoost model successfully identifies 6078 out of the total 7477 positive cases. This outcome underscores the potential use of our method in the healthcare system, demonstrating its ability to accurately identify a substantial proportion of positive cases.

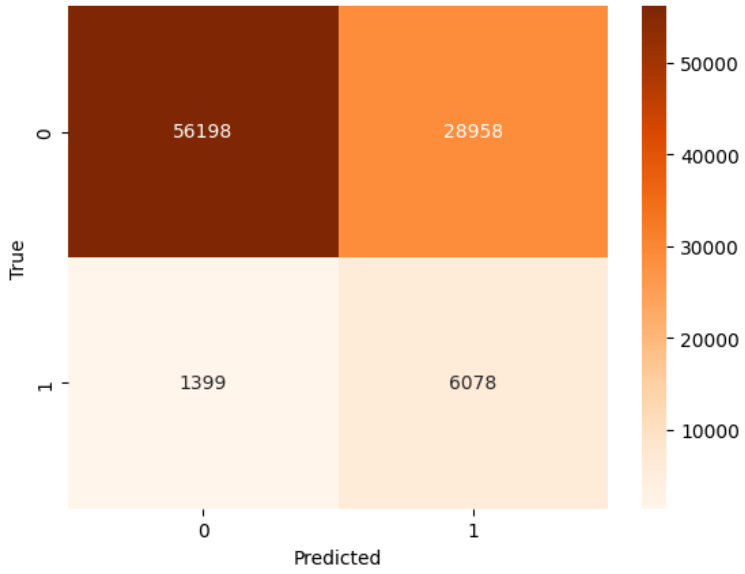

Figure 17: Confusion matrix of XGBoost&SMOTE.

6.1.3 Results interpretation with ADASYN

Distinguishing itself from SMOTE, ADASYN introduces an adaptive element to the synthetic data generation process. While SMOTE uniformly augments the minority class with synthetic instances, ADASYN takes a dynamic approach. It concentrates its synthetic efforts on regions of the feature space where minority instances are scarce, providing a more fine-tuned adjustment to the data landscape.

Illustrated in Figure 18, ADASYN dynamically resamples the dataset, increasing the minority class instances to 200721, a considerable augmentation, while it leaves the majority class untouched, preserving its original count at 198647 instances.

```
from imblearn.over_sampling import ADASYN

print("Before OverSampling- counts of label '1': {}".format(sum(y_train==1)))
print("Before OverSampling- counts of label '0': {} \n".format(sum(y_train==0)))

adasyn = ADASYN()
X_resampled, y_resampled = adasyn.fit_resample(X_train,y_train)

print("After ReSampling with Adasyn - '1': {}".format(sum(y_resampled==1)))
print("After ReSampling with Adasyn - '0': {}".format(sum(y_resampled==0)))
```

```
Before OverSampling- counts of label '1': 17494
Before OverSampling- counts of label '0': 198647

After ReSampling with Adasyn - '1': 200721
After ReSampling with Adasyn - '0': 198647
```

Figure 18: Over sampling with ADASYN.

The impact of ADASYN on model performance can be seen in Table 7. Examining the performance metrics in this table, it is discerned that ADASYN, while bolstering the recall of the models, exhibits a slightly more limited effect compared to its oversampling counterpart, SMOTE. Specifically Logistic Regression and Random Forest, showcase again noteworthy improvement, achieving a commendable 55% on recall, while Decision Trees and Boosting algorithms, demonstrate increased accuracy but grapple with a much lower recall.

Table 7: Performance results after implementing ADASYN.

Model	***Accuracy / Training Accuracy***	***Recall / Training Recall***	***Precision/ Training Precision***	***F1-score/ Training F1-score***	***AUC/ Training AUC***
Logistic Regression	0.73 / 0.74	0.56 / 0.56	0.16 / 0.16	0.25 / 0.25	0.73 / 0.73
Decision Tree	0.84 / 1.00	0.26 / 1.00	0.17 / 1.00	0.20 / 1.00	0.57 / 1.00
Random Forest	0.73 / 0.74	0.56 / 0.56	0.16 / 0.16	0.25 / 0.26	0.78 / 1.00
Gradient Boosting	0.83 / 0.83	0.40 / 0.40	0.21 / 0.21	0.27 / 0.28	0.77 / 0.77
XGBoost	0.90 / 0.92	0.11 / 0.15	0.32 / 0.44	0.16 / 0.23	0.80 / 0.84
CatBoost	0.92 / 0.92	0.08 / 0.14	0.40 / 0.63	0.13 / 0.23	0.81 / 0.86

Furthermore, after also examining the models' performance on the training data, a significant disparity in the results becomes evident for Decision Trees when compared to their performance on the testing data. The model achieves exceptionally high accuracy on the training data but performs poorly on unseen data. This discrepancy raises concerns about potential overfitting, suggesting that the model may have captured noise or specific patterns that do not generalize effectively beyond the training set.

We proceed with the optimization of our models, utilizing the hyperparameters recommended by the GridSearchCV and Optuna algorithms. The outcomes of this optimization are detailed in Table 8. Notably, XGBoost and CatBoost maintain a commendable AUC rate of 80%, while concurrently enhancing their sensitivity to 81% and 80%, respectively, surpassing the SMOTE's performance on the same models. Also, despite the improvement in F1-score compared to previous results, there is still room for enhancement.

Table 8: Performance results after optimization (ADASYN).

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.75 / 0.76	0.50 / 0.50	0.16 / 0.16	0.25 / 0.25	0.71 / 0.72
Decision Tree	0.85 / 0.98	0.23 / 0.78	0.18 / 0.88	0.20 / 0.82	0.59 / 1.00
Random Forest	0.74 / 0.76	0.56 / 0.50	0.16 / 0.16	0.25 / 0.25	0.78 / 1.00
Gradient Boosting	0.83 / 0.90	0.41 / 0.17	0.21 / 0.27	0.27 / 0.21	0.77 / 0.79
XGBoost	0.67 / 0.68	0.81 / 0.89	0.17 / 0.18	0.28 / 0.31	0.80 / 0.85
CatBoost	0.70 / 0.71	0.80 / 0.90	0.18 / 0.20	0.30 / 0.34	0.81 / 0.87

6.1.4 Results interpretation with SMOTE-Tomek

Following, we employed the SMOTE-Tomek algorithm, which is a hybrid resampling method, commonly used to handle the imbalance in the data. With this process, as seen in Figure 19, the minority class is augmented into 198031 instances, offering a substantial reinforcement to its representation within the dataset, while simultaneously, the Tomek Links algorithm identifies and eliminates instances that form Tomek Links - pairs of instances of different classes that are closest to each other, facilitating a focused reduction in the majority class.

```
from imblearn.combine import SMOTETomek
from imblearn.under_sampling import TomekLinks

print("Before OverSampling- counts of label '1': {}".format(sum(y_train==1)))
print("Before OverSampling- counts of label '0': {} \n".format(sum(y_train==0)))

smtomek = SMOTETomek(sampling_strategy = 1.0, n_jobs = -1, random_state = 42)
X_resampled, y_resampled = smtomek.fit_resample(X_train,y_train)

print("After OverSampling with SMOTE-Tomek - '1': {}".format(sum(y_resampled==1)))
print("After OverSampling with SMOTE-Tomek- '0': {}".format(sum(y_resampled==0)))
```

```
Before OverSampling- counts of label '1': 17494
Before OverSampling- counts of label '0': 198647

After OverSampling with SMOTE-Tomek - '1': 198031
After OverSampling with SMOTE-Tomek- '0': 198031
```

Figure 19: Hybrid resampling with SMOTE-Tomek.

We then apply our machine learning models, and the results can be seen in Table 9.

Table 9: Performance results after implementing SMOTE-Tomek.

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.70 / 0.70	0.63 / 0.61	0.16 / 0.16	0.25 / 0.25	0.74 / 0.74
Decision Tree	0.84 / 1.00	0.27 / 1.00	0.17 / 1.00	0.21 / 1.00	0.58 / 1.00
Random Forest	0.70 / 0.70	0.63 / 0.61	0.16 / 0.16	0.25 / 0.25	0.78 / 1.00
Gradient Boosting	0.83 / 0.84	0.40 / 0.40	0.21 / 0.22	0.28 / 0.28	0.77 / 0.78
XGBoost	0.90 / 0.91	0.12 / 0.16	0.31 / 0.43	0.17 / 0.24	0.80 / 0.84
CatBoost	0.91 / 0.92	0.08 / 0.15	0.40 / 0.63	0.14 / 0.24	0.81 / 0.87

Now that our data is appropriately balanced, we find considerable satisfaction in the 91% accuracy achieved by boosting algorithms. Nevertheless, the 8% recall rate indicates a significant limitation, as the models tend to categorize all observations into the majority class, masking their ability to effectively identify positive cases. In light of this, we favour the performance of Logistic Regression and Random Forests, striking a balance between relatively high accuracy and a satisfactory level of recall.

Following the fine-tuning of the models, as seen in Table 10, we see that unfortunately Decision Trees overfit to the training data, Logistic Regression and Random Forest drop their performance, but interestingly, the noteworthy performance of CatBoost and XGBoost emerges. Their impressive consistency between the testing and training sets underscores its exceptional generalization ability. CatBoost not only attains an impressive 70% accuracy but also excels in identifying positive cases, achieving an 81% recall rate. Likewise, XGBoost outperforms all models and achieves an 82% in its sensitivity, which is the highest result until now. These results underline the effectiveness and robustness of boosting algorithms in handling the intricacies of the dataset.

Table 10: Performance results after optimization (SMOTE-Tomek).

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.76 / 0.76	0.49 / 0.50	0.16 / 0.17	0.25 / 0.25	0.72 / 0.73
Decision Tree	0.86 / 0.98	0.21 / 0.79	0.20 / 0.93	0.20 / 0.85	0.59 / 1.00
Random Forest	0.76 / 0.76	0.49 / 0.50	0.16 / 0.17	0.25 / 0.25	0.78 / 1.00
Gradient Boosting	0.83 / 0.84	0.40 / 0.40	0.21 / 0.22	0.28 / 0.28	0.77 / 0.78
XGBoost	0.67 / 0.68	0.82 / 0.89	0.17 / 0.19	0.28 / 0.31	0.80 / 0.85
CatBoost	0.70 / 0.71	0.81 / 0.90	0.18 / 0.21	0.30 / 0.34	0.81 / 0.87

6.1.5 Results interpretation with SMOTE-ENN

Finally, we apply the SMOTE-ENN hybrid resampling algorithm on the BRFSS data, which combines synthetic data generation and data refinement. This method strategically augments the minority class to 193714 instances, significantly bolstering its representation in the dataset, while concurrently, it undertakes a pruning of the majority class, reducing its observations to 133109 from the original count of 198647, as illustrated in Figure 20.

```
from imblearn.combine import SMOTEENN

print("Before OverSampling- counts of label '1': {}".format(sum(y_train==1)))
print("Before OverSampling- counts of label '0': {} \n".format(sum(y_train==0)))

sm = SMOTEENN()
X_resampled, y_resampled = sm.fit_resample(X_train,y_train)

print("After OverSampling with SMOTEENN - '1': {}".format(sum(y_resampled==1)))
print("After OverSampling with SMOTEENN - '0': {}".format(sum(y_resampled==0)))
```

```
Before OverSampling- counts of label '1': 17494
Before OverSampling- counts of label '0': 198647

After OverSampling with SMOTEENN - '1': 193714
After OverSampling with SMOTEENN - '0': 133109
```

Figure 20: Hybrid resampling with SMOTE-ENN.

The performance results after applying the SMOTE-ENN algorithm are presented in Table 11.

Table 11: Performance results after implementing SMOTE-ENN.

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.60 / 0.61	0.79 / 0.80	0.14 / 0.15	0.24 / 0.25	0.75 / 0.76
Decision Tree	0.79 / 0.91	0.42 / 1.00	0.17 / 0.48	0.24 / 0.65	0.62 / 0.95
Random Forest	0.60 / 0.61	0.79 / 0.80	0.14 / 0.48	0.24 / 0.25	0.79 / 0.99
Gradient Boosting	0.74 / 0.74	0.67 / 0.68	0.19 / 0.19	0.29 / 0.30	0.78 / 0.79
XGBoost	0.85 / 0.86	0.42 / 0.50	0.25 / 0.30	0.31 / 0.38	0.80 / 0.84
CatBoost	0.87 / 0.89	0.37 / 0.48	0.28 / 0.36	0.32 / 0.41	0.81 / 0.86

Upon examining these findings, a noteworthy improvement is evident across all models in terms of recall, while also a quite good accuracy rate is preserved, following the application of the SMOTE-ENN technique. For instance, Gradient Boosting exhibits a substantial increase from 4% to 67% without any additional optimization. Additionally, it is remarkable to note the impressive 79% recall achieved by both Logistic Regression and Random Forest. Additionally, with our dataset now balanced, we can delve into the AUC metric. It reveals that nearly all our models exhibit a commendable ability to correctly classify instances, with Random Forest notably achieving an impressive 80%.

Proceeding to the optimization of the models, we carefully selected hyperparameters tailored to our data, giving priority to those assigning different weights to each class. This strategic choice creates a heightened focus on the minority class, which is particularly significant in our case. The results, as depicted in Table 12, are promising. Notably, we avoided overfitting across all models, as evidenced by the similar performance on both the training and test sets. Moreover, we attained the peak performance for each model. CatBoost, in particular, surpassed expectations with a remarkable 88% recall, coupled with a decent accuracy rate and an impressive 82% AUC rate.

Table 12: Performance results after optimization (SMOTE-ENN).

Model	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Logistic Regression	0.66 / 0.66	0.71 / 0.73	0.15 / 0.16	0.25 / 0.26	0.74 / 0.75
Decision Tree	0.79 / 0.91	0.41 / 0.99	0.17 / 0.48	0.24 / 0.65	0.62 / 0.95
Random Forest	0.66 / 0.66	0.71 / 0.73	0.15 / 0.16	0.25 / 0.26	0.78 / 0.98
Gradient Boosting	0.74 / 0.74	0.67 / 0.68	0.19 / 0.19	0.29 / 0.30	0.78 / 0.79
XGBoost	0.61 / 0.61	0.87 / 0.94	0.15 / 0.17	0.26 / 0.28	0.80 / 0.85
CatBoost	0.63 / 0.63	0.88 / 0.94	0.16 / 0.17	0.27 / 0.29	0.82 / 0.86

This outcome underscores the efficacy of combining CatBoost with the hybrid SMOTE-ENN algorithm, especially for healthcare practitioners dealing with real-life imbalanced data. As illustrated in Figure 21, this combination proves to be highly beneficial, enabling the identification of a substantial number of positive cases—an advantageous outcome for the healthcare system.

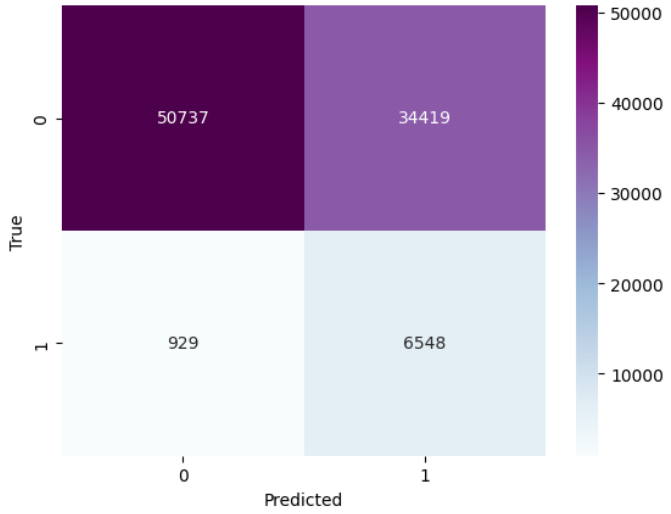

Figure 21: Confusion matrix of the peak performance (CatBoost&SMOTE-ENN).

6.2 Deep Learning implementation

The inherent complexity and non-linearity of Artificial Neural Networks (ANNs) have led to their adoption in tasks that demand the extraction of intricate patterns and representations from complex data types, and therefore has traditionally been

synonymous with unstructured data, such as images, audio, and natural language [58]. The deterministic nature of structured data seemingly contradicts the flexibility and adaptability that ANNs offer. Nevertheless, in our exploration we will showcase how ANNs, when properly harnessed, can surpass the performance of traditional machine learning algorithms on structured data.

Model architecture

The foundation of our model lies in the sequential arrangement of densely connected layers. Each layer is tailored to capture distinct features and patterns from the input data, transforming them into increasingly abstract representations. The architecture details of our proposed ANN can be seen in Figure 22.

Input Layer

The initial layer consists of 128 units, each equipped with a Rectified Linear Unit (ReLU) activation function. ReLU is chosen for its ability to introduce non-linearity into the model, allowing it to learn complex relationships within the input data [59]. The number of units is determined manually based on the complexity of the dataset and the need for the network to extract diverse features.

Dropout Layer (Regularization)

Following the first dense layer, a dropout layer with a dropout rate of 0.5 is introduced. Dropout is a regularization technique that randomly drops a fraction of the connections during training, preventing the model from relying too heavily on specific features and enhancing its generalization capabilities [60]. The chosen dropout rate strikes a balance between mitigating overfitting and retaining valuable information.

Second Dense Layer

The second dense layer further refines the learned features with 64 units and a ReLU activation function. The reduced number of units in this layer allows for a gradual transition from the expansive feature space captured by the initial layer to more concise and informative representations.

Second Dropout Layer

To fortify the model against overfitting, a second dropout layer is inserted with the same dropout rate of 0.5. The sequential arrangement of dropout layers serves as a protective mechanism, encouraging the network to learn robust and transferable features [60].

Output Layer

The final layer, comprising a single unit and a sigmoid activation function, transforms the learned features into a probability score. The sigmoid activation is apt for binary classification tasks, producing a probability indicating the likelihood of the presence of a cardiovascular disease.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               2944

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 11265 (44.00 KB)
Trainable params: 11265 (44.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 22: Architecture details of proposed ANN model.

Class imbalance

To address the class imbalance, present in our dataset, except for the resampling algorithms we put into use, we assign a class weight to the positive class during model compilation, emphasizing on the importance of correctly predicting cases at high risk for CVDs. On the raw data, the class weight for the positive class is set to 10, while on the methods using resampling the class weight for the positive class is set to 2.

Loss Function

The model is compiled using the Adam optimizer, a popular choice for its efficiency in updating weights during training. Network output was then compared to the desired output, and error was calculated using binary cross-entropy loss, as shown in (1):

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \quad [22](1)$$

Early stopping

In our ANN training, the implementation of early stopping is executed through the EarlyStopping callback provided by the Keras library. The callback is configured to monitor the 'val_loss', the loss on the validation set. Training is halted if the validation loss fails to improve over a predefined number of epochs, known as the patience parameter, which we set to 10. This effectively prevents the model from excessively tailoring its parameters to the training data, ensuring a more generalized and robust model.

Prediction and evaluation

Once trained, we employ our model to make predictions on the training and the testing sets, and the predicted probabilities are thresholded at 0.5 to obtain binary predictions. The model's performance when employed with each of the resampling algorithms is presented in Table 13.

Table 13: Performance results for the ANN.

ANN	*Accuracy/ Training Accuracy*	*Recall/ Training Recall*	*Precision/ Training Precision*	*F1-score/ Training F1-score*	*AUC/ Training AUC*
Raw data	0.74 / 0.73	0.78 / 0.80	0.20 / 0.21	0.32 / 0.33	0.76 / 0.77
SMOTE	0.58 / 0.78	0.80 / 0.93	0.14 / 0.71	0.23 / 0.81	0.68 / 0.77
ADASYN	0.63 / 0.77	0.74 / 0.92	0.15 / 0.70	0.24 / 0.80	0.68 / 0.76
SMOTE-Tomek	0.62 / 0.77	0.77 / 0.94	0.15 / 0.70	0.24 / 0.80	0.69 / 0.77
SMOTE-ENN	0.54 / 0.83	0.87 / 0.96	0.14 / 0.79	0.24 / 0.87	0.69 / 0.80

The results taken, reveal several noteworthy observations. Firstly, our successful handling of overfitting stands out as a significant achievement, particularly in the context of applying Artificial Neural Networks (ANN) to structured, non-complex data—a major challenge in such scenarios.

Upon closer examination, it becomes apparent that, on the raw data, our ANN surpasses the performance of all previously employed machine learning (ML) models. Impressively, it attains a recall rate of 78%, while matching the accuracy rate achieved by the top-performing ML model at 74%. This accomplishment underscores the efficacy of our ANN in extracting meaningful patterns from the data.

Notably, the combination of our ANN and the SMOTE-ENN hybrid resampling algorithm yields exceptional results, reaching its peak recall rate of 87%. This synergy demonstrates the effectiveness of incorporating data resampling techniques to enhance the performance of our ANN.

While our ANN, coupled with the SMOTE, ADASYN, and SMOTE-Tomek algorithms, falls short of outperforming our optimized boosting algorithms, it is noteworthy that it achieves a performance close to theirs. This implies that, even in scenarios where

boosting algorithms maintain a slight edge, our ANN remains a competitive and promising rival.

The nuanced interplay between our ANN and various resampling techniques showcases its adaptability and potential to deliver robust performance across diverse data types.

7 Discussion

Our investigation into predicting cardiovascular diseases on a real-life dataset, through both machine learning and deep learning algorithms, unfolds critical findings explained in this comprehensive evaluation. We employed 6 ML models, Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, XGBoost, and CatBoost on our raw data, and we uncovered initial challenges tied to dataset imbalance. Despite achieving high overall accuracy, our models struggled to effectively identify positive cases, having a notably suboptimal recall rate of only 4%. This discrepancy raised concerns about the models' sensitivity in detecting individuals with heart-related conditions, prompting an in-depth investigation into mitigating bias and enhancing the models' performance in identifying positive cases.

Our optimization strategy involved a detailed hyperparameter tuning process using GridSearchCV for Logistic Regression, Decision Trees and Random Forests and Optuna for Gradient Boosting and CatBoost. XGBoost demonstrated the importance of adaptive approaches, outperforming grid searches with manually found hyperparameters.

In the subsequent phase of our investigation, we focus on a resampling process, applying various oversampling and hybrid sampling algorithms to achieve class balance. Initially, two oversampling algorithms, SMOTE and ADASYN, were implemented. SMOTE showcased a remarkable improvement in all models' performance, especially when coupled with XGBoost, achieving 67% accuracy and an impressive 80% recall rate. ADASYN, while not initially surpassing SMOTE, exhibited enhanced performance after fine-tuning, with XGBoost achieving an 81% recall rate with the same as previously accuracy rate.

Moving forward, we explored two hybrid resampling algorithms, SMOTE-Tomek and SMOTE-ENN, strategically combining under-sampling and oversampling methods. While SMOTE-Tomek faced challenges with overfitting in certain models, it proved effective with boosting algorithms. XGBoost achieved an 82% recall rate, and CatBoost reached 81%. SMOTE-ENN demonstrated immediate improvements in the models' generalization ability and effective detection of positive cases. The combination of SMOTE-ENN and CatBoost showcased our study's peak performance, achieving an 88% recall rate, indicating a minimal miss rate for positive cases in the dataset.

Finally, we demonstrated the efficacy of Artificial Neural Networks (ANN) for structured data when handled appropriately. The proposed ANN, coupled with the SMOTE-ENN algorithm, achieved an 87% recall rate and 70% on the AUC metric, and on the raw data it outperformed all our ML models even when optimized, affirming its potential as a valuable tool for healthcare practitioners dealing with imbalanced data.

In Figure 23, we present our recommended optimal resampling algorithm–predictive model combinations, providing a comprehensive overview of their achieved recall rates.

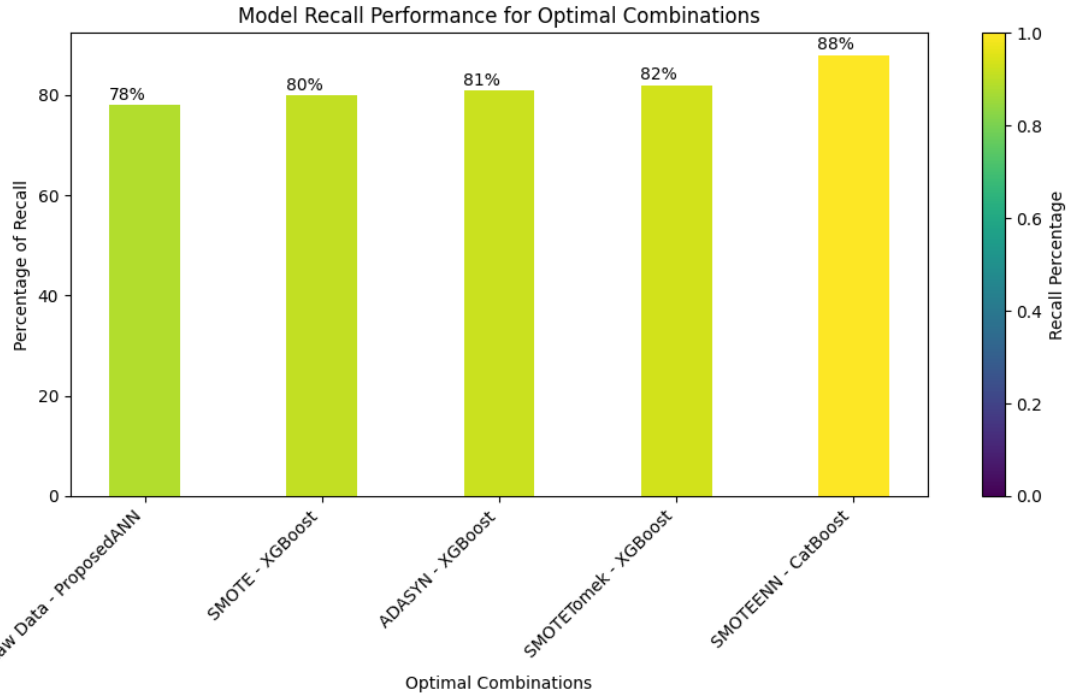

Figure 23: Model Recall Performance for Optimal Combinations.

8 Conclusion and Future Work

This study effectively tackled the challenge of class imbalance within a real-life dataset. By systematically comparing Machine Learning (ML) and Deep Learning (DL) algorithms, coupled with various resampling techniques, we discerned the optimal combination that maximizes the recall metric.

8.1 Conclusion

In conclusion, our journey through the prediction of cardiovascular diseases using a real-life dataset, explored through both machine learning and deep learning algorithms, has demonstrated critical insights crucial for healthcare practitioners and researchers alike. The evaluation results presented in the preceding chapter underscore the nuanced challenges inherent in dealing with imbalanced datasets and the profound implications for the accurate identification of positive cases, when their representation is limited in the dataset.

Our initial findings revealed a trade-off between high accuracy and poor recall for positive cases, mostly because of a major imbalance present, prompting a meticulous investigation into addressing this challenge. Our optimization strategies, encompassing hyperparameter tuning and hybrid resampling techniques, served as pivotal interventions to enhance model performance and sensitivity. Five different experiments were conducted separately for each of the resampling algorithms to find the best-performing model. We implemented two over-sampling and two hybrid resampling, algorithms, to address the imbalance in the dataset and used 6 machine learning models as well as an artificial neural network for our predictions. These interventions showcased marked improvements, with optimized boosting algorithms along with hybrid resampling methods achieving impressive recall rates, with CatBoost specifically coupled with SMOTE-ENN achieving a remarkable 88%. Furthermore, the integration of Artificial Neural Networks (ANN) into our study demonstrated their capacity to excel in handling structured imbalanced data, offering an additional layer of robustness in detecting positive cases in the healthcare sector.

Moreover, it is imperative to underscore the indispensable role of proper electronic health records (EHR) in excelling predictive analytics for healthcare. The wealth of

structured and unstructured data within EHR has proven instrumental in constructing a comprehensive analysis of patient health profiles. The information encapsulated in EHR not only enhances the predictive capabilities of our models but also facilitates a holistic comprehension of individual health trajectories, laying the groundwork for a robust prevention system.

As we conclude this chapter of our research, it becomes evident that the synergy between advanced analytics, innovative algorithms, and the depth of information within EHR is the cornerstone of transformative breakthroughs in predictive healthcare.

8.2 Future Work

There are several promising avenues to explore when building upon the foundations laid in our study. Firstly, as mentioned earlier, the quality of the EHR can have a significant impact on the performance of the proposed models. The integration of more diverse and extensive datasets could enhance the generalizability of predictive models and improve their ability to identify patients at high-risk. Additionally, investigating the impact of incorporating genetic and biomarker data would provide a more holistic understanding of cardiovascular disease risk, as there is existing evidence that strongly suggests that there is a potential correlation.

Exploring advanced machine learning techniques, and deep learning architectures, the application of transfer learning could significantly enhance the predictive capabilities of our models. By leveraging knowledge gained from pre-trained models on structured data, on health-related domains, we can potentially improve the performance of CVD risk prediction. Also, the exploration of explainable AI methods could address the interpretability challenges associated with complex models, fostering greater trust and adoption in clinical domains and assist even further, health practitioners, understand the reasons behind the models' decision making.

Bibliography

[1] Cardiovascular Diseases (Cvds), "World health organization," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Cleveland Clinic, 2022. Cardiovascular Disease. Retrieved August 01, 2023, from https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

[3] M. De Hert, J. Detraux, and D. Vancampfort. (2022) "The intriguing relationship between coronary heart disease and mental disorders," Dialogues in Clinical Neuroscience, vol. 20, DOI: 10.31887/DCNS.2018.20.1/mdehert

[4] Heart Foundation (n.d.). Common medical tests to diagnose heart conditions. Retrieved August 05, 2023, from https://www.healthywa.wa.gov.au/Articles/A_E/Common-medical-tests-to-diagnose-heart-conditions

[5] University of Rochester Medical Center Rochester, 2023. Tests to diagnose heart problems. Retrieved August 05, 2023, from

https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=85&contentid=P00208

[6] Mayo Clinic, 2022. Heart disease. Retrieved August 05, 2023, from https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124

[7] Wendy Wisner ,2023. What Is Preventive Health and Why Is It Important? Retrieved on August 10, 2023, from https://www.healthline.com/health/what-is-preventive-health-and-why-is-it-important

[8] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. J Big Data 9, 3 (2022). https://doi.org/10.1186/s40537-021-00553-4

[9] EIT Health, (2020). Early diagnostics: shaping healthcare and society through new technologies. https://eithealth.eu/wp-content/uploads/2020/09/EIT-Health-paper_Early-Diagnostics_Shaping-Healthcare-Society.pdf

[10] Foresee Medical (n.d.). Predictive analytics in healthcare. Retrieved August 07, 2023, from https://www.foreseemed.com/predictive-analytics-in-healthcare

[11] Mary K. Pratt, 2021. Predictive analytics in healthcare: 12 valuable use cases.

Retrieved August 07, 2023, from

https://www.techtarget.com/searchbusinessanalytics/tip/Predictive-analytics-in-healthcare-12-valuable-use-cases

[12] Nadejda Alkhaldi, 2022. Predictive analytics in healthcare: 7 ways to save time and money. Retrieved on August 10, 2023, from https://itrexgroup.com/blog/predictive-analytics-in-healthcare-top-use-cases/

[13] B. R. Lindman, S. V. Arnold, R. Bagur et al. (2020) "Priorities for patient-centered research in valvular heart disease: a report from the national heart, lung, and blood institute working group" *Journal of American Heart Association*, vol. 9, no. 9, Article ID e015975.

[14] NHS (n.d.). Heart Failure. Retrieved August 01, 2023, from https://www.nhs.uk/conditions/heart-failure/

[15] Cleveland Clinic, 2022. Arrhythmia. Retrieved August 01, 2023, from https://my.clevelandclinic.org/health/diseases/16749-arrhythmia

[16] Cleveland Clinic, 2022. Heart Valve Disease. Retrieved August 01, 2023, from https://my.clevelandclinic.org/health/diseases/17639-what-you-need-to-know-heart-valve-disease

[17] R. J. Hinchliffe, J. R. W. Brownrigg, J. Apelqvist et al. (2016) "IWGDF guidance on the diagnosis, prognosis and management of peripheral artery disease in patients with foot ulcers in diabetes" *Diabetes*, vol. 32, pp. 37–44, https://doi.org/10.1002/dmrr.2698

[18] Cleveland Clinic, 2022. Deep Vein Thrombosis. Retrieved August 01, 2023, from https://my.clevelandclinic.org/health/diseases/16911-deep-vein-thrombosis-dvt

[19] MAYOCLINIC, (2022). Myocarditis. https://www.mayoclinic.org/diseases-conditions/myocarditis/symptoms-causes/syc-20352539

[20] Asif, Daniyal, Mairaj Bibi, Muhammad Shoaib Arif, and Aiman Mukheimer. 2023. "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization" Algorithms 16, no. 6: 308. https://doi.org/10.3390/a16060308

[21] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration", *International Journal of Intelligent Systems and Applications*, vol. 7, no. 12, pp. 75–82, Nov. 2015, doi: 10.5815/ijisa.2015.12.08.

[22] Sarra, Raniya & Dinar, Ahmed & Mohammed, Mazin. (2023). "Enhanced accuracy for heart disease prediction using artificial neural network", *Indonesian Journal of Electrical Engineering and Computer Science*, 29. 375-383. 10.11591/ijeecs.v29.i1.pp375-383.

[23] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146

[24] Farhat Ullah, Xin Chen, Khairan Rajab, Mana Saleh Al Reshan, Asadullah Shaikh, Muhammad Abul Hassan, Muhammad Rizwan, Monika Davidekova, "An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Predication", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 1906466, 12 pages, 2022. https://doi.org/10.1155/2022/1906466

[25] Tick V. K., Meeng N. Y., Mohammad N. F., Harun N. H., Alquran H., Mohsin M. F. M. (2021, August). "Classification of Heart Disease using Artificial Neural Network", *Journal of Physics*: Conference Series (Vol. 1997, No. 1, p. 012022). IOP Publishing. [26] Weng, SF, Reps, J, Kai, J, Garibaldi, JM, and Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. (2017) 12:e0174944. doi: 10.1371/journal.pone.0174944

[27] Abdullah Alqahtani, Shtwai Alsubai, Mohemmed Sha, Lucia Vilcekova, Talha Javed, "Cardiovascular Disease Detection using Ensemble Learning", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 5267498, 9 pages, 2022. https://doi.org/10.1155/2022/5267498

[28] Trigka M. , Dritsas E., Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models. Sensors 2023, 23, 1193. https://doi.org/10.3390/s23031193

[29] Mirza Muntasir Nishat, Fahim Faisal, Ishrak Jahan Ratul, Abdullah Al-Monsur, Abrar Mohammad Ar-Rafi, Sarker Mohammad Nasrullah, Md Taslim Reza, Md Rezaul Hoque Khan, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset", *Scientific Programming*, vol. 2022, Article ID 3649406, 17 pages, 2022. https://doi.org/10.1155/2022/3649406

[30] T. R. Mahesh, V. Dhilip Kumar, V. Vinoth Kumar, Junaid Asghar, Oana Geman, G. Arulkumaran, N. Arun, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9005278, 11 pages, 2022. https://doi.org/10.1155/2022/9005278

[31] Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton, "An efficient convolutional neural network for coronary heart disease prediction", Expert Systems with Applications, Volume 159, 2020, 113408, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113408.

[32] A. Gupta, R. Kumar, H. Singh Arora and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," in IEEE Access, vol. 8, pp. 14659-14674, 2020, doi: 10.1109/ACCESS.2019.2962755.

[33] G. Paragliola and A. Coronato, "An hybrid ECG-based deep network for the early identification of high-risk to major cardiovascular events for hypertension patients," *Journal of Biomedical Informatics*, vol. 113, Article ID 103648, 2021.

[34] Mohammed Nasir Uddin, Rajib Kumar Halder (2021). “An ensemble method based multi-layer dynamic system to predict cardiovascular disease using machine learning approach”, *Informatics in Medicine Unlocked*, Volume 24, 100584, ISSN 2352-9148

[35] Ahmed Al Ahdal, Manik Rakhra, Rahul R. Rajendran, Farrukh Arslan, Moaiad Ahmad Khder, Binit Patel, Balaji Ramkumar Rajagopal, Rituraj Jain, "Monitoring Cardiovascular Problems in Heart Patients Using Machine Learning", *Journal of Healthcare Engineering*, vol. 2023, Article ID 9738123, 15 pages, 2023. https://doi.org/10.1155/2023/9738123

[36] Syed Javeed Pasha, E. Syed Mohamed, “Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction*”, Informatics in Medicine Unlocked*, Volume 32, 2022, 101064, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2022.101064.

[37] S. Sharma and M. Parmar, “Heart diseases prediction using deep learning neural network model,” International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 3, pp. 2244–2248, 2020.

[38] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 8387680, 11 pages, 2021. https://doi.org/10.1155/2021/8387680

[39] Subramani Sivakannan, Varshney Neeraj, Anand M. Vijay, Soudagar Manzoore Elahi M., Al-keridis Lamya Ahmed, Upadhyay Tarun Kumar, Alshammari Nawaf, Saeed Mohd, Subramanian Kumaran, Anbarasu Krishnan, Rohini Karunakaran (2023). “Cardiovascular diseases prediction by machine learning incorporation with deep learning”, *Frontiers in Medicine*, Volume 10, ISSN 2296-858X

[40] Sumaira Ahmed, Salahuddin Shaikh, Farwa Ikram, Muhammad Fayaz, Hathal Salamah Alwageed, Faheem Khan, Fawwad Hassan Jaskani, "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models*", Journal of Sensors*, vol. 2022, Article ID 3730303, 21 pages, 2022. https://doi.org/10.1155/2022/3730303

[41] Nousi, C., Belogianni, P., Koukaras, P., Tjortjis, C. (2022). Mining Data to Deal with Epidemics: Case Studies to Demonstrate Real World AI Applications. In: Lim, CP., Vaidya, A., Jain, K., Mahorkar, V.U., Jain, L.C. (eds) Handbook of Artificial Intelligence in Healthcare. Intelligent Systems Reference Library, vol 211. Springer, Cham. https://doi.org/10.1007/978-3-030-79161-2_12

[42] Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, Eman M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method", *Complexity*, vol. 2021, Article ID 6663455, 10 pages, 2021. https://doi.org/10.1155/2021/6663455

[43] Center for disease control (2022), 2021 BRFSS Survey Data and Documentation, https://www.cdc.gov/brfss/annual_data/annual_2021.html

[44] National Center for Chronic Disease Prevention and Health Promotion (2022), The Nation's Risk Factors and CDC's Response,

https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

[45] Teboul A. (2020), "Building predictive models for Heart disease", https://www.linkedin.com/pulse/building-predictive-models-heart-disease-alex-teboul

[46] Hao-Yun Hsieh, Chang-Fu Su, Shu-I Chiu (2022), "Constructing Multiple Layers of Machine Learning for the Early Detection of Cardiovascular Diseases", EasyChair

[47] Lupague R.M.J.M., Mabborang R.C., Bansil A.G., Lupague M.M. (2023) Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors, *European Journal of Computer Science and Information Technology*, 11 (3), 44-58

[48] Centers for Disease Control and Prevention (2022), "Adult Body Mass Index" , https://www.cdc.gov/obesity/basics/adult-defining.html

[49] Swastik S. (2023), "SMOTE for Imbalanced Classification with Python", https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[50] Raden A. (2021), "Imbalanced Classification in Python: SMOTE-ENN Method", https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50

[51] Hairani H., Anggrawan A., Priyanto D. (2023) Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link, *Int. J. Inform. Visualization*, 7(1) - March 2023 258-264

[52] Ahmed, Intisar, "A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING AND DATA MINING" (2022). Electronic Theses, Projects, and Dissertations. 1591

[53] N.Permatasari, Shafiyah A.S., A.L.Irfiansyahand M. G.Al-Haqqoni (2022). PREDICTING DIABETES MELLITUS USING CATBOOST CLASSIFIER AND SHAPLEY ADDITIVE EXPLANATION (SHAP) APPROACH", BAREKENG: J. Il. Mat. & Ter., vol. 16, iss. 2, pp. 615-624, June, 2022.

[54] Rohit K. Chowdary, Bhargav P., Nikhil N., Varun K., Jayanthi D.(2022) "Early heart disease prediction using ensemble learning techniques", *Journal of Physics : Conference Series*, volume 2325 012051, DOI 10.1088/1742-6596/2325/1/012051

[55] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama. (2019) "Optuna: A Next-generation Hyperparameter Optimization Framework" , arXiv, https://doi.org/10.48550/arXiv.1907.10902

[56] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, “Classification models for heart disease prediction using feature selection and PCA”, *Informatics in Medicine Unlocked*, Volume 19, 2020, 100330, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100330.

[57] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, Qian Wang, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", Computational and Mathematical Methods in Medicine, vol. 2017, Article ID 8272091, 11 pages, 2017. https://doi.org/10.1155/2017/8272091

[58] Neeraja Vaidya, “How Artificial Neural Networks Unlock Insights from Unstructured Data”, https://blog.aureusanalytics.com/blog/how-artificial-neural-networks-unlock-insights-from-unstructured-data

[59] Bharath Krishnamurthy (2022), “An Introduction to the ReLU Activation Function”, https://builtin.com/machine-learning/relu-activation-function

[60] Database Camp (2023), “What is the Dropout Layer?”, https://databasecamp.de/en/ml/dropout-layer-en