T.C.

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DIAGNOSIS OF CARDIOVASCULAR (CVD) DISEASE USING HIGH PERFORMANCE MACHINE LEARNING SYSTEM

Mustafa FAYEZ

Doctor of Philosophy

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, 2021

DIAGNOSIS OF CARDIOVASCULAR (CVD) DISEASE USING HIGH PERFORMANCE MACHINE LEARNING SYSTEM

by

Mustafa Adil Fayez Fayez

Electrical and Computer Engineering

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2021

The dissertation titled "DIAGNOSIS OF CARDIOVASCULAR (CVD) DISEASE USING HIGH PERFORMANCE MACHIN LEARNING SYSTEM" prepared by "Mustafa Adil Fayez FAYEZ" and submitted on (29 / 04 / 2021) has been accepted unanimously for the degree of PhD in Electrical and Computer Engineering.

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Sefer KURNAZ	School of Engineering and Natural Sciences, Altinbas University	_____
Prof. Dr. Osman Nuri UÇAN	School of Engineering and Natural Sciences, Altinbas University	_____
Prof. Dr. Mesut RAZBONYALI	School of Engineering and Natural Sciences, Maltepe University	_____
Asst. Prof. Dr. Zeynep ALTAN	School of Engineering and Architecture, Beykent University	_____
Asst. Prof. Dr. Oguz KARAN	School of Engineering and Natural Sciences, Altinbas University	_____

I hereby declare that this dissertation meets all format and submission requirements of a PhD dissertation.

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Mustafa Adil Fayez FAYEZ

Signature

Accepted by Altınbaş University Institute of Graduate
Studies on the following date:
29 / 04 / 2021

DEDICATION

I dedicated this dissertation to Allah and my great supervisor, Asst. Prof. Dr. Sefer Kurnaz, as well as to my family, who supported me during my studies, and to my homeland, Iraq.

ACKNOWLEDGEMENTS

First and foremost, I'd like to convey my heartfelt appreciation to Asst. Prof. Dr. Sefer Kurnaz, my PhD counselor, for his unwavering guidance, compassion, inspiration, passion, and vast expertise. Throughout the study and writing of this thesis, his advice was invaluable. For my PhD study, I couldn't have asked for a better advisor and tutor.

ABSTRACT

DIAGNOSIS OF CARDIOVASCULAR (CVD) DISEASE USING HIGH PERFORMANCE MACHINE LEARNING SYSTEM

FAYEZ, Mustafa Adil Fayez,

PhD, Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: March / 2021

Pages: (82)

One of the most prevalent illnesses in middle-aged people is heart failure. Cardiovascular disease (CAD) is a common coronary disease with a high mortality risk of the various forms of heart disease. Medical imaging, such as angiography, is the most effective technique for diagnosing CAD. Angiography, on the other hand, is notorious for being pricey and causing a variety of side effects. In the first part of our work, we will address one of the most sophisticated and effective systems in the medical sector, specifically cardiac diagnostics, in this thesis. This advanced methodology, which is split into two sections, was developed with the aid of the best programming method, Python. To start, feature engineering and preprocessing with a high-performance re-sampling system and a neighborhood cleaning rule is used (NCL). Second, hyper-parameter optimization, API advanced features, and super learner models are examples of advanced and optimization machine learning models. We have used streamlined SVM to construct a high-performing diagnostic model. We built advanced architectures for high-

performance machine learning, such as AutoML, advanced XGBoost, and advanced ensemble bagging models, in the second part of our study. We obtained the best performance using Auto-ML and advanced SMOTE method, which had an accuracy of 87%, advanced deep learning had an accuracy of 80%, and advanced bagging models had an accuracy of 82 percent.in addition, we achieved 86% with XGBoost and 92% used stacking model. Finally, we assume that our advances can change the way doctors view machine learning using advanced and high-performance machine learning tools, as well as promote the wider usage of Artificial Intelligence (AI) techniques in clinical settings.

Keywords: CVD disease, NCL, Advanced machine learning models, Optimization, API function advanced, Super learner, Advanced Bagging models, Advanced XGBoost, AutoML, Python.

ÖZET

YÜKSEK PERFORMANSLI MAKİNE ÖĞRENME SİSTEMİ KULLANILARAK KARDİYOVASKÜLER (CVD) HASTALIĞININ TANI

Mustafa Adil Fayez, FAYEZ
PhD, Elektrik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi
Danışman: Yrd. Dr. Sefer Kurnaz
Tarih: Mart / 2021
Sayfalar: (82)

Orta yaşlı insanlarda en yaygın hastalıklardan biri kalp yetmezliğidir. Kardiyovasküler hastalık (CAD), çeşitli kalp hastalıklarının mortalite riski yüksek olan yaygın bir koroner hastalıktır. Anjiyografi gibi tıbbi görüntüleme, CAD tanısı için en etkili tekniktir. Öte yandan anjiyografi, pahalı olması ve çeşitli yan etkilere neden olmasıyla ünlüdür. Çalışmalarımızın ilk bölümünde, tıp sektöründeki en sofistike ve etkili sistemlerden birini, özellikle kardiyak tanıyı bu tezde ele alacağız. İki bölüme ayrılan bu gelişmiş metodoloji, en iyi programlama yöntemi python yardımıyla geliştirilmiştir. Başlamak için, yüksek performanslı bir yeniden örnekleme sistemi ve mahalle temizleme kuralı (NCL) ile mühendislik ve ön işleme özelliği kullanılır. İkinci olarak, hiper parametre optimizasyonu, API gelişmiş özellikleri ve süper öğrenci modelleri gelişmiş ve optimizasyon makine öğrenimi modellerine örnektir. Yüksek performanslı bir tanı modeli oluşturmak için aerodinamik SVM kullandık. Çalışmamızın ikinci bölümünde AutoML, gelişmiş XGBoost ve gelişmiş topluluk torbalama modelleri gibi yüksek performanslı makine öğrenimi için gelişmiş mimariler inşa ettik. %87 doğruluk oranına sahip Auto-ML ve gelişmiş SMOTE yöntemini kullanarak en iyi performansı elde ettik, gelişmiş derin öğrenme %80 doğruluk

oranına sahipti ve gelişmiş torbalama modelleri 82 percent.in ilave isabet oranına sahipti, XGBoost ve %92 kullanılmış istifleme modeli ile %86'ya ulaştık. Son olarak, ilerlemelerimizin doktorların gelişmiş ve yüksek performanslı makine öğrenimi araçlarını kullanarak makine öğrenimine bakış açısını değiştirebileceğini ve klinik ortamlarda Yapay Zeka (AI) tekniklerinin daha geniş kullanımını teşvik edebileceğini varsayıyoruz.

Anahtar Kelimeler: CVD hastalığı, NCL, Gelişmiş makine öğrenimi modelleri, Optimizasyon, Gelişmiş API işlevi, Süper öğrenci, Gelişmiş Torbalama modelleri, Gelişmiş XGBoost, AutoML, Python.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Cases

CHAPTER 1

1.1 INTRODUCTION

A category of conditions that arise in the blood arteries and heart are cardiovascular diseases (CVDs). This covers a variety of illnesses, including (CAD) disease, commonly known as (CHD) disease, which is the world's most serious heart disease. The key screening technique for the irregular shrinking of heart arteries is angiography. Researchers are seeking to identify new diagnosis approaches because of their cost and risks. In 2035, it is expected that the number of Americans with CVD would increase to 131.2 million-45% of the overall U.S. population. Past assessments also determined that 100 million Americans would have cardiovascular disease by 2030. Instead, in 2015, the nation surpassed the benchmark and is already on target with more than 131 coronary disease patients by 2035 [1].

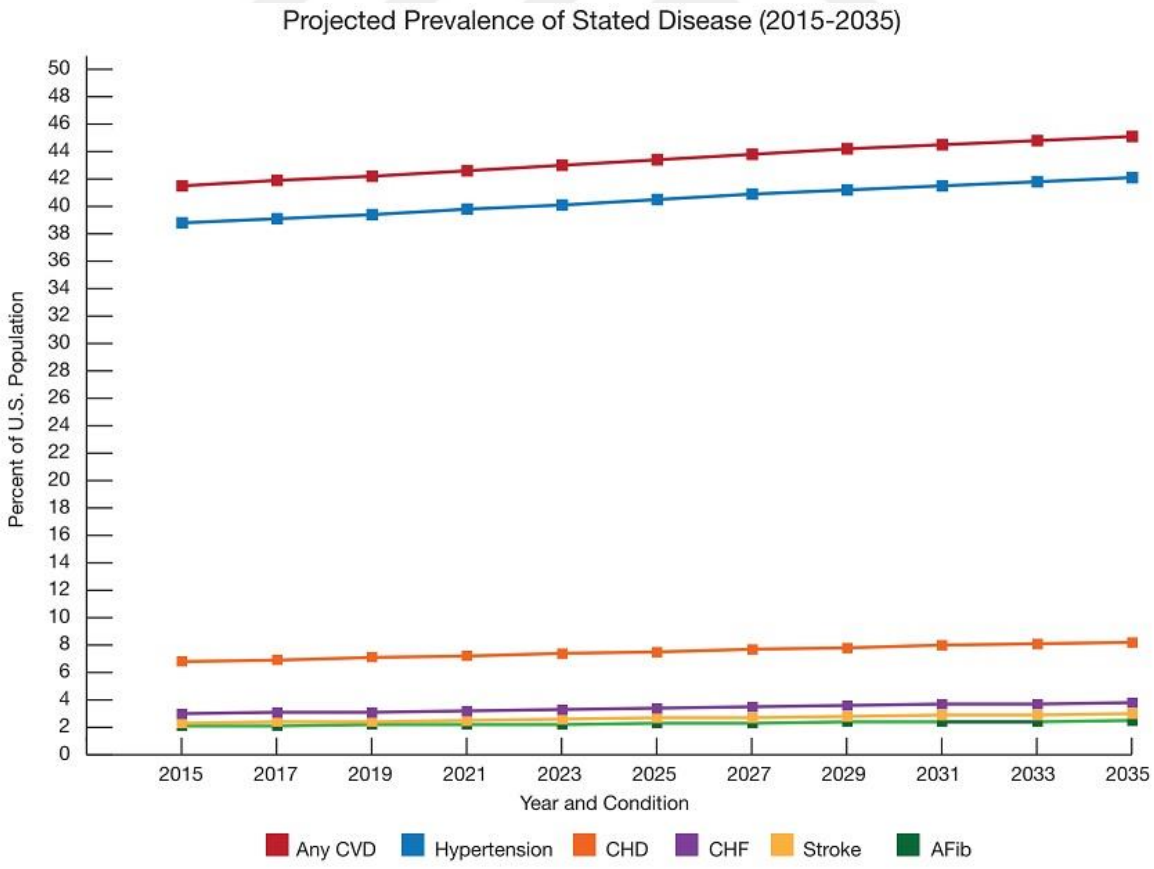

Figure 1. 1: Projected Prevalence of CVD Disease (2015-2035)[2]**.**

In the last five years, machine learning has seen significant development. The usage of data processing strategies for Machine Learning (ML) will alleviate the need for human experience and the risk of human error while increasing the precision of prediction [3]. Cantered on learned relationships between variables in the input dataset, ML algorithms implement flexible prediction models. This will avoid fixed diagnostic templates, such as the AHA recommendations, from being oversimplified. Artificial intelligence calculations by health analysts also improved vital excitement. For example, an on-going analysis indicates that some interactions addressed with medical care use or plan to use sophisticated machine learning models in imagery [4]. The emergence of Auto-ML is technically ground-breaking in the medicinal and therapeutic domains. In their practice and research, Auto-ML can allow it by eliminating the high technical barrier; doctors can use AI methods more commonly. To search for the right AI model setup, Auto-ML can be seen as the end-to-end approach on an arbitrarily specified dataset [5]. Each specification is the outcome of several decisions about which algorithm, optimization type, or hyper-parameter to be using. Because of the massive amount of structures in the quest space, it is computationally expensive to find the correct model. Due to modern devices such as (GPU) and (TPU) units, and extremely strong search systems, Auto-ML techniques have been capable to scale up dramatically as computer processing technology progresses. Recent research has shown that computer-generated classifiers have matched or even exceeded those formed by human authorities [6]. The goal is to automate the process of creating an advanced and optimized model that provides any provided dataset with competitive performance. This entails data pre-processing automation, extraction of functions, tuning of hyper-parameters, and choice of algorithms [7]. This thesis would shed light on an imperative and emerging ML method for doctors and biomedical sciences. Our results will reveal insight into what an automated machine learning technology will offer, how easy it is to use the software, and how well it works compared to the human experience [8]. When checking our method for cardiovascular data, other types of biomedical data are required to maintain the findings. In the past, a number of experiments have been performed to get ML systems to identify patterns in the data to take early cardiovascular disease detection into consideration [9]. The goal of this thesis is to use an advanced ML mode as a resolution to the directly above declared glitches and efforts to determine the performance and time advantages that advanced ML systems and optimization would have to give over a professionally designed methodology.

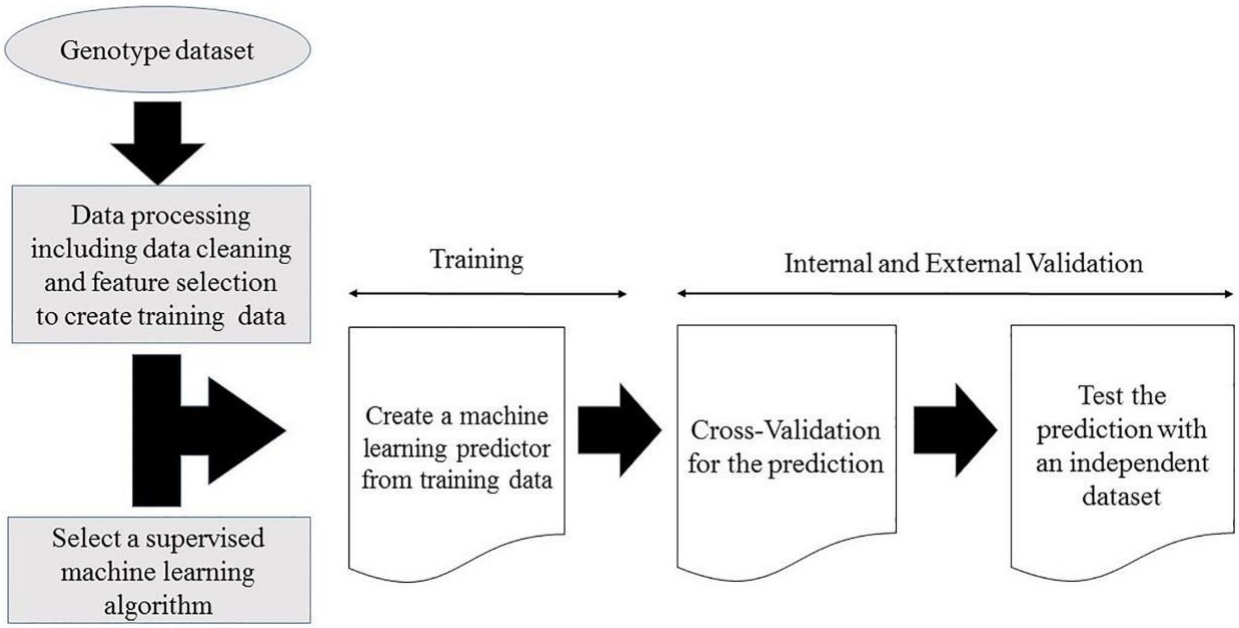

Figure 1. 2: Machine Learning System Structure [10]**.**

1.2 PROBLEM STATEMENT

Heart disease has been deemed the most important cause of death globally for several years. In order to complete the diagnosis of CVD for patients, health experts utilize their expertise and abilities and usually, most diagnostic evidence obtained from patients is held in archives or directories. For users, these large numbers of unorganized health registries make little sense. Machine learning is a classification tool that we can use by remembering past cases and recycling details and experience to solve this problem. In such instances, these approaches translate knowledge into valuable information that can further support the CVD diagnostic mechanism. This method is designed to offer the greatest potential contribution to the diagnosis of medical data by doctors and clinical experts. Therefore, in defining and assessing the health condition of people, the device will support doctors and health experts. The key purpose of this study is to build a therapeutic dataset method to assess the status of patients, which can save time and resources and bring benefit to doctors' quality options, which in turn prefer patients to heal which can be implemented when there is a lack of specialists. This research aims to help decide which characteristics are more relevant for the diagnosis of CVD and which machine learning system is more useful for the development of a CVD diagnostic framework.

1.3 THE CONTRIBUTION

Our study provides the following contributions:-

- For the first time, we will use high performance of pre-processing methods for cardiovascular CVD dataset including re-sampling method neighbourhood cleaning rule (NCL) and (SMOTE) over-sampling systems to decrease indistinguishable instances that contribute to imbalanced data problems becoming solved.
- We are evaluating the use of advanced deep learning including hyper-parameter optimization model, advanced super learner, and model of optimizing SVM before and after re-sampling process to develop advanced cardiovascular disease diagnosis system.
- We test the usage of the high-performance method including AutoML, advanced XGBoost utilizing parameter tuning model and advanced Bagging model after resampling process.
- We compare our advanced system performance with those of another system with extensive machine learning and computer programming tool.
- On CVD dataset we have detailed experimental results.

1.4 THESIS STRUCTURE

This work is covered in a variety of papers in journals, either written or planned for publication. (Chapters 1-5):

Chapter 1 this section includes a general introduction. As the state-of-the-art approaches, it comes with industry and research needs and towards machine learning, concentrating on the problems and possibilities for medical data and classification through advanced artificial intelligence technologies also the problem statement and our contribution.

Chapter 2 provides a concise review and background of the optimization and machine learning based methods in medical data classification, and discusses the previous study comparison against this work and diagnosis problem.

Chapter 3 presents the methodology of optimization classifier and advanced model structure with the description of CVD dataset and the pre-processing methods using high performance model including the explanation of resampling process.

Chapter 4 presents the results before and after resampling method and discussion with details using different types of tables and figures.

Chapter 5 presents a conclusion of our study, future work and the references.

CHAPTER 2

2.1 INTRODUCTION

Coronary heart disorders (CVD) have the largest prevalence of non-communicable diseases, contributing to death in most countries around the world, from the study of a variety of science, technological and medical outlets. Therefore, in order to minimize management expenses, it is important to produce a diagnostic model or framework for CVD. Prevention is, therefore, necessary. Today, several innovative studies have suggested approaches and applications that use AI and ML strategies to forecast and diagnose CVD using data mining algorithms.

2.2 CVD DISEASE

A variety of cardiac and blood vessel disorders are expressed in cardiovascular disease. The incidence and treatment expenses of elevated specifically includes heart rate, coronary artery disease (CAD), heart problems (CHF), strokes, atrial fibrillation (AFib), among other heart diseases from now until 2035. CVD heart failure arises when it is weakened or diseased by the main blood vessels that feed the heart with blood and oxygen [11]. Cardiovascular disorder (CVD), coronary heart disease (CVD), and ischemic heart disease are three primary causes of heart disease. In certain instances, CVD occurs and causes death in multiple countries, particularly for males and females of various ages in the United States and China. Plaque deposits in the arteries of the heart are typically the cause of CVD. The following figure illustrates the shape of CVD and its impact on the heart's arteries.

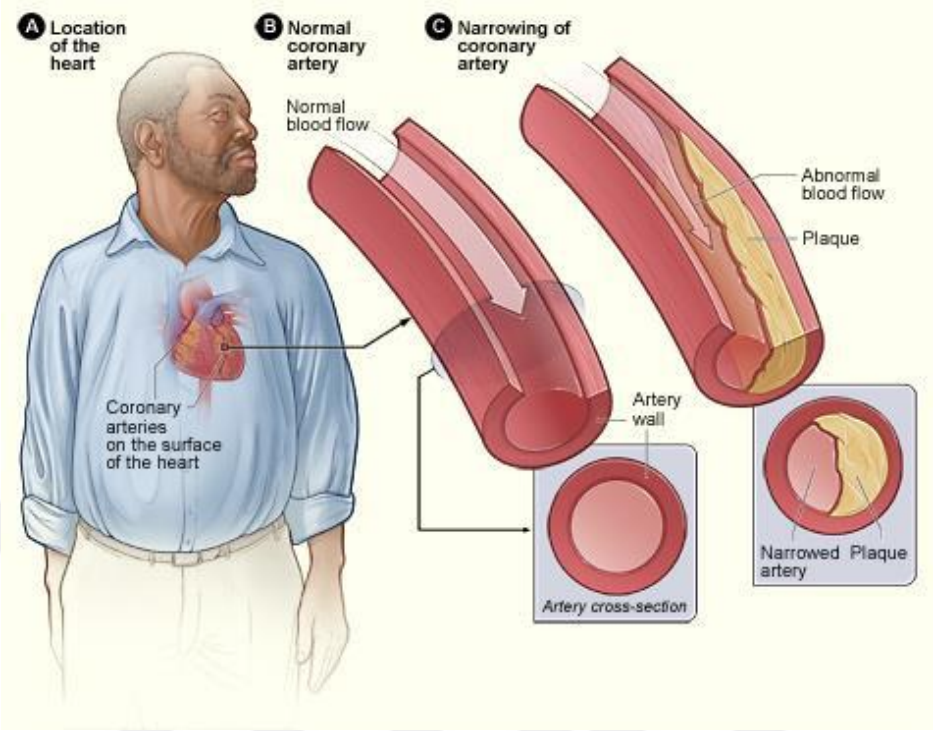

Figure 2. 1: CVD Disease Description [12].

A heart attack may be caused by total blockages, and smoking, elevated blood pressure, high cholesterol, and a sedentary lifestyle are the major factors of this condition.

2.3 DIAGNOSIS PROBLEM

An accurate health review, which involves a study of the patient's personal and family medical records as well as measures of blood pressure, cholesterol, and physical fitness, is a significant factor in avoiding CVD [13]. This could be time-intensive and expensive. Machine learning has long been explored in the field of healthcare by the use of ML instruments and equipment to facilitate the production of data processing of broad and complicated medical datasets. In the detection, prediction, and deep comprehension of health care datasets, an accepted or implemented high performance ML system on the drug side is of great importance. In order to promote and prevent any errors in hospitals with early detection of illness and prevention inpatient fatalities, the purposes of these submissions require action center evaluations. We will reduce time and decrease costs with high performance ML system, thereby rendering the

diagnostic problem simpler by constructing a model to find the best approach for diagnosing CVD disease with high precision.

2.4 BACKGROUND OF CVD WITH MACHINE LEARNING

CVD diagnosis and study is of special importance and consideration to the community because of the great loss of human life among individuals, especially in European countries. Providing proper medical identification and the successful delivery of therapies is the primary concern of health care organizations. Bad treatment options end in extreme and inappropriate consequences. It is important to promote a clinical decision-making mechanism in the health care sector with more modern technologies, including a computer-based information system. A group of physical and classification methods that can be used to uncover secret facts and expertise from a patient dataset are part of ML research [14]. This information is an additional pool of expertise for healthcare practitioners to make wise clinical choices, which in turn enhances patient protection, increases the quality of life of patients, and improves patient detection performance. There are several programs built by numerous researchers to facilitate the diagnosis of CVD disease by experts and physicians. There are several machine learning algorithms that are applied to diagnosing CVD and used in ML techniques, such as kernel density, neural networks, bagging algorithms, sequential minimum optimization, direct kernel self-organizing maps, and SVM. To aid in the diagnostic procedures, ML classification algorithms are used.

2.5 MACHINE LEARNING

Machine learning, a common sub discipline of AI, describes different strategies by defining interaction trends between variables to solve complicated problems with big data. Machine learning, in comparison to conventional statistics, focuses on the development of advanced clinical decision systems that enable doctors render predictions more precise rather than simplistic calculated score systems. It is possible to categorize machine learning into 3 forms of learning: supervised; unsupervised; and strengthening [15]. Algorithms use a dataset labeled by human beings in supervised learning to predict the expected and proven effect. For classification and regression problems, supervised learning is fantastic, but it involves a lot of information is timewasting so humans have to mark the details. Unmonitored learning, on the other side, attempts to recognize novel pathways of illness, genotypes or phenotypes from secret trends present in the results. The aim of unsupervised learning is to discover secret trends in the data without human input. For example, it may be called guided learning to instruct medicinal inhabitants (labels) before they see patients. In addition, it may be considered unsupervised learning to encourage medical residents to see patients (no labels), then allow them to learn from their errors to derive up with their particular ideas (optimize). To optimize the precision of automatic prediction, certain systems, such as (ANN), may be educated utilizing supervised or unsupervised learning. Finally, it is possible to interpret reinforcement learning as a combination between guided and unsupervised learning. In the choosing of a curve, learning curves and (AUC) are major factors [16]. Machine-learning algorithm, whereas in the option of conventional data-processing methods, c-statistics are essential as with conventional statistics, in order to improve its output on the training dataset until testing, machine learning needs adequate The correct algorithms and learning databases (also defined as sample size in standard statistics). In conventional statistics, alpha does not reach 0.05, while there should be no under fitting (suboptimal data) or over fitting in machine learning (noisy data) [17]. Over fitting also happens because, compared to the size of the training dataset (i.e. so many parameters), a model is unnecessarily complicated, and might not be true for the testing dataset. In comparison, CVD in men, CVD deprived of neuromuscular dysplasia with CVD may theoretically be exempt from AI. This topic is related to under Data fitting.

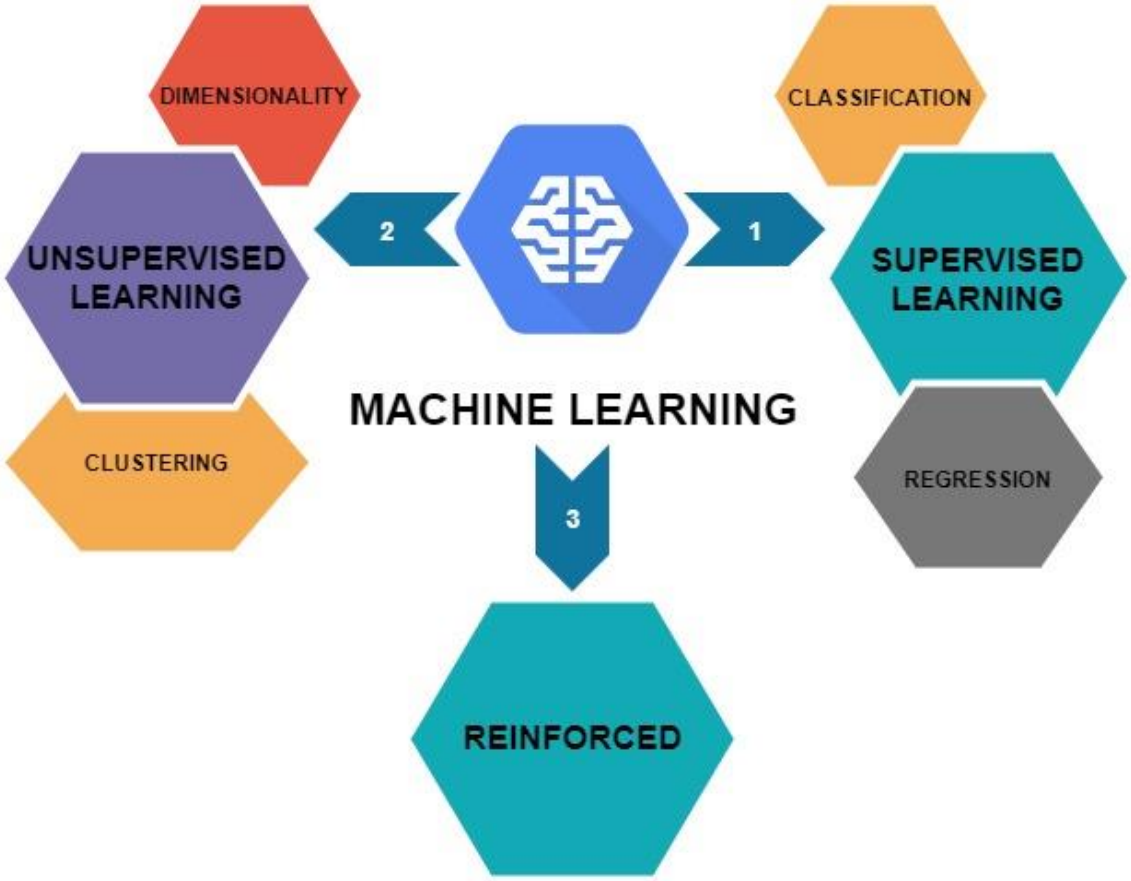

Figure 2. 2: Machine Learning Structure.

2.6 SUPERVISED LEARNING

You train the model on a labeled dataset in supervised learning, ensuring we have both raw input data and its outcomes. We separate our knowledge into a training dataset and a test dataset where our network is conditioned by the training dataset, whilst the test dataset serves as fresh data to forecast outcomes or to see the consistency of our model [18]. Therefore, our model learns from the outcomes shown in guided instruction, the same way a teacher teaches his students when the teacher already understands the findings. In guided learning, consistency is what we accomplish, as model perfection is typically high. The model operates easily because, since we already have ideal outcomes in our dataset, the training period required is smaller. Without even understanding a prior goal, this model forecasts reliable effects on unseen data or fresh data. In some guided learning models, we reverse the output result and understand further in order to achieve the highest possible accuracy [19]. The value may discretely present a group with each input instance and predicted value associates, or it can be actual or continuous value. The algorithm knows the input patterns that produce the expected performance, and it can now be used to predict the correct output of a never-seen input until the algorithm is qualified.

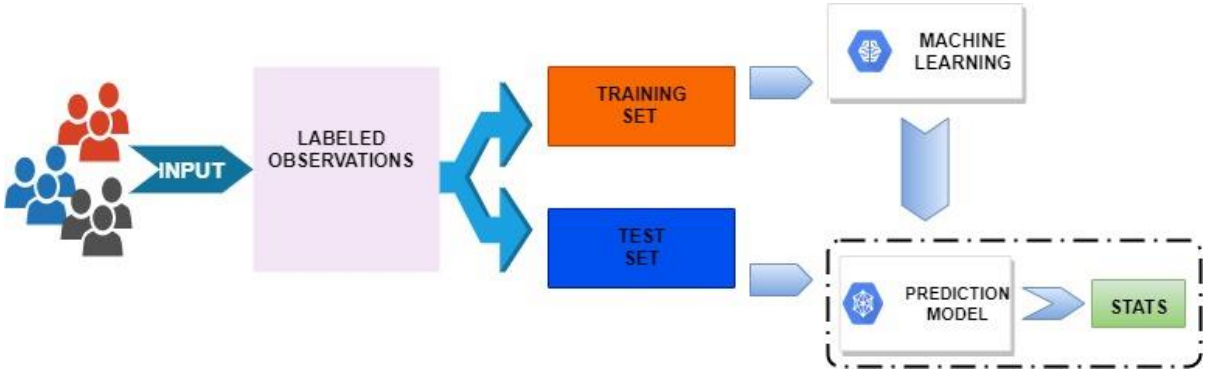

Figure 2. 3: Supervised Learning Structure.

2.7 LIMITATIONS OF SUPERVISED LEARNING

Tiny training datasets may contribute to erroneous judgments in evaluating datasets in supervised learning where the training data are biased; As a result, required data are used to train the model, as well as other sets of data to verify it [19]. Supervised learning often involves manually tagging the training datasets to model only established performance outcomes. In addition, while a given training dataset can match several functions for which supervised learning subsequently makes decisions about the "right" function that matches the knowledge, it can create bias that directs the learning procedure to choose one supposition or feature over another.

2.8 UNSUPERVISED LEARNING

The data used to practice is neither categorized nor named in the dataset in unsupervised learning. Unsupervised learning experiments about how structures can infer from unlabeled data a feature to define a secret framework. Seeking trends in the data is the key challenge of unsupervised learning [20]. If a model learns to create trends, it can easily predict patterns in the form of clusters for every new dataset. The machine does not work out the correct performance, but it examines the details and can draw data set inferences to explain secret constructs from unlabeled data.

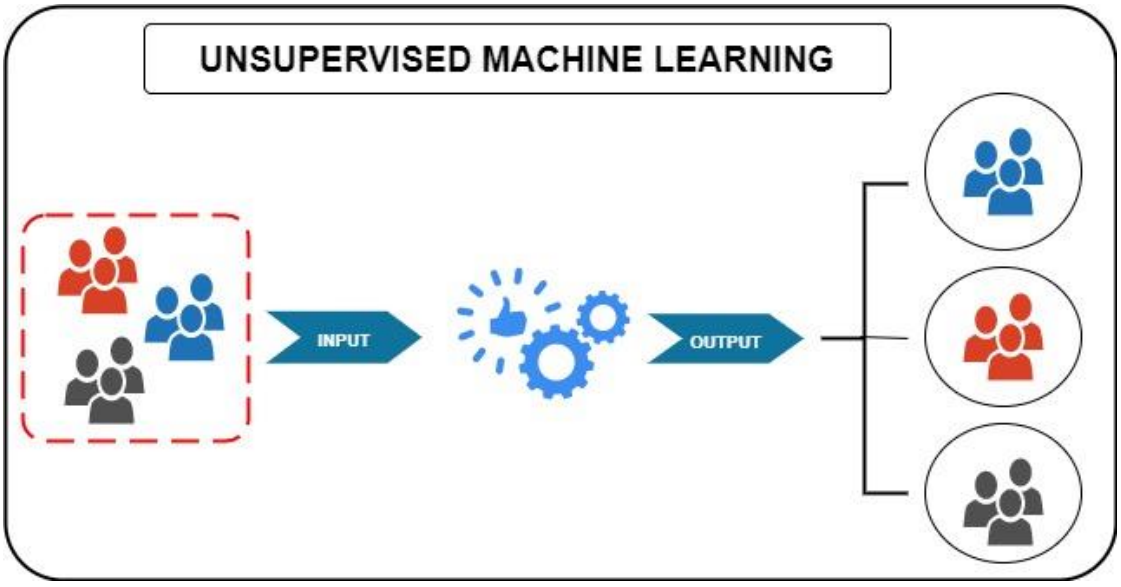

Figure 2. 4: Unsupervised Learning Structure.

2.9 LIMITATIONS OF UNSUPERVISED LEARNING

Difficulty recognizing the original cluster pattern, which could theoretically contribute to prejudices, is one significant drawback in unsupervised learning. This could result in incorrect judgments since the final cluster pattern relies on the original cluster pattern. In multiple cohorts, therefore, it requires confirmation [21]. In addition, certain complicated glitches lead to restrictions that are not easily solved deprived of supervised training; therefore, to find the optimal algorithm, it can involve manually labeled data. Noisy data, such as 3D-STE photos, for instance, must be denoted manually before predictions can be accurate. Unstructured data can therefore include manual hand coding in some components and unmonitored methodologies in others, and succeeding evaluation for the best performance.

2.10 DEEP LEARNING

Deep learning uses the workings of the human brain across several layers of ANN that can produce automatic feedback forecasts (teaching data). It has been a hot topic in AI since it seems exciting and is a rising area. It can be very useful in image recognition (e.g., Facebook facial recognition, Google image search) [22]. They may also be prepared for unsupervised learning activities in an unsupervised way. There is, in comparison, no restriction on working memory. It also operates well on a variety of noisy specifics, such as strain photography and 3DSTE results. The usage of artificial real-time CV imaging with greater three-dimensional and chronological precision would be enabled by deep learning algorithms, eventually enhancing the standard of treatment and reducing costs.

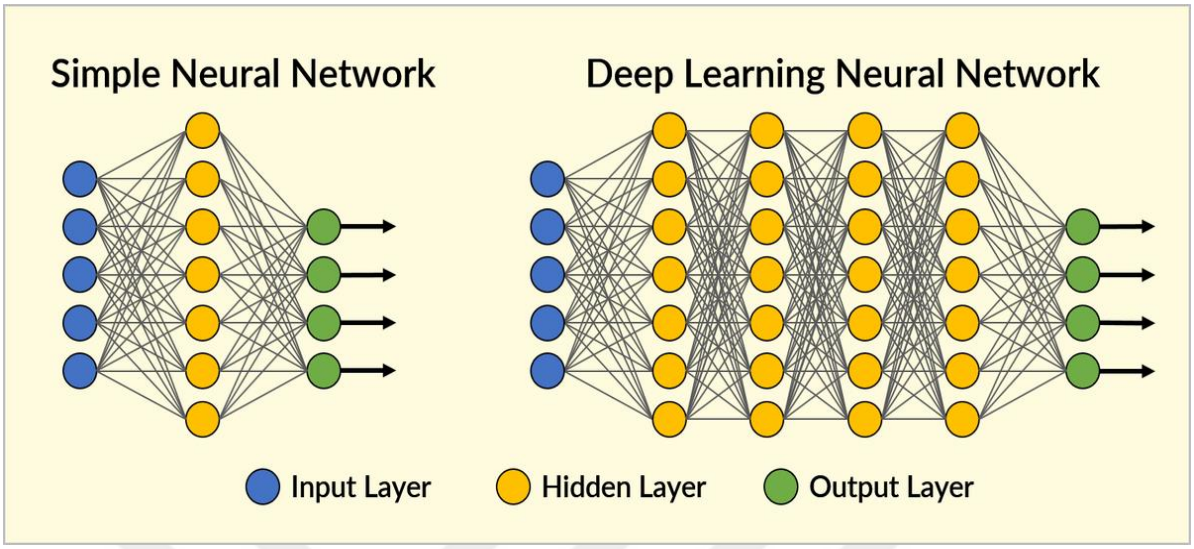

Figure 2. 5: Deep Learning Structure [23].

2.11 LIMITATIONS OF DEEP LEARNING

Deep learning is a form of nonlinear processing that has many parameters and layers; over fitting may also be a serious obstacle that may contribute to low predictive results. It can help to avert over-fitting by growing the size of the training data or declining the amount of secret layers. A broad training dataset is also needed for deep learning, which involves cooperation between institutions and EHR linkages [24]. Deep-learning research often includes deep-learning-capable computers, such as unit-accelerated computation for graphics processing, such as NVIDIA and Amazon EC2. It's often time consuming to build and utilizing deep learning technique. Finally, although deep learning with multiple layers would not improve precision, it will lengthen the training time.

2.12 DEEP KNOWLEDGE

Deep intelligence uses several levels of (ANN) to produce automated predictions from a training data collection, simulating the capabilities of the human brain. In order to distinguish pictures (cardiac angiography and magnetic resonance) this information may be commonly utilized [25]. It may also be equipped, such as a drug-drug relationship, to conduct an unsupervised learning activity. In comparison, there are no limitations on occupied memory. It has proven to be preferable to other ML methods, such as SVM, since, relative to the two levels of MVR, it can use several layers and transformations. Since a non-linear study with several parameters and many layers is typically deep knowledge, over fitting may be high, leading to poor predictive results.

CHAPTER 3

3.1 RELATED WORK

In this section we will discuss the latest and most recent scientific work in this field compared to the current study by developing in the system in both sides of the processing of the tested data and the amount of significant improvement in the results of the highly efficient advanced machine learning system.

Chayakrit Krittanawong et al.[26] Proposed system of AI for detailed cardiovascular medicine. This paper offered insight into the use of AI in medical treatment for cardiovascular (CVD) and explores its possible function in promoting the accuracy of (CVD).

Xiao Liu et al.[27] Based on a classification scheme of hybrid cardiac disease diagnostic tools based on the (RFRS) model. Centered on the C4.5 classifier, according to a cross-validation technique for the jackknife, an ensemble classifier is suggested to achieve a max classification accuracy of 92.59 percent. The findings validate that the efficiency of the proposed method is superior to that of the classification strategies previously stated.

A systematic analysis and evaluation of strategies for ML in the area of heart disease proposed by Seyedamin Pouriyeh et al.[28]. The aim of this paper is to use Ensemble Machine Learning Techniques to study and evaluate the accuracy of different data mining classification systems for the prediction of heart disease. 10-Fold Cross-Validation was used to increase the amount of data, which would have been limited otherwise.

Syed Muhammad Saqlain Shah et al. they proposed Feature Extraction via the Key Component of Simultaneous Probabilistic Analysis and Detection of Heart Failure [29]. In the current projection, the suggested approach extracts high-impact characteristics by utilizing Key Feature Analysis Probabilistic (PCA). In contrast to the current research demonstrating its effect, the statistical findings obtained by the proposed methodology are discussed. The suggested approach obtained a precision of 82.18 percent, 85.82 percent, and 91.30 percent for the data collection in Cleveland, Hungary, and Switzerland, respectively.

Identification of significant features and data mining strategies suggested by Mohammad Shafenoor Amin et al.[30]. In forecasting heart disease this study aims to recognize essential features and strategies for data mining that can enhance the precision of CVD prediction. Using numerous combinations of features and seven classification techniques, prediction models were created. The results of the experiment show that the model of heart disease prediction created using the substantial characteristics discovered and the best-performing data mining technique (i.e. Vote) achieves an accuracy of 87.4 percent in heart disease prediction.

Rubens Moura Campos Zeron and Carlos Vicente Serrano Junior [26] developed an artificial intelligence for the diagnosis CVD disease. This study provides a snapshot of the use of AI in clinical cardiovascular treatment and explores its possible role in promoting cardiovascular precision medicine.

Nabaouia Louridi et al.[31] Have based their studies on the identification of CVD disorders using ML in 2019. In this article, to increase the efficiency of CVD disease prediction, they used a successful pre-processing tool. This helps to diagnose the cardiac condition of a sufferer and helps a specialist to properly recognize whether or not a person has CVD. Using SVM with a linear kernel, the 86.8 percent precision of the score was achieved.

They introduced a deep learning method to the classification of CVD disease by updated ECG signals from decomposition in empirical mode with Nahian Ibn Hasan et al.[32]. The model suggested achieved a combined precision of 97.70 percent, 99.71 percent, and 98.24 percent respectively.

In 2019, Sivaranjani.R and Dr.N.Yuvaraj [33] operated on an AI device with a multilayer perceptron to anticipate earlier cardiac characteristics and functionality. This thesis aims to establish a medical predictor method to forecast issues with the heart. The Deep Learning-based predictive model is intended for use in diagnosis. The best activation function is selected from the highest precision bar map when a procedure's output is evaluated for three activation functions.

Meghana Padmanabhan et al. [34] focused on doctor-accessible machine learning in 2019: a cardiovascular disease risk prediction case report. They looked at the expected incidence of cardiovascular disease. They allow a comparison of (auto-ML) methodology with a graduate student utilizing crucial measures to prove their arguments, along with the entire period needed to shape (ML) models and the inference of classification accuracies on unseen test data sets. They were 74 percent right with 70,000 cardiovascular disease results, and 85 percent with Heart UCI data set.

Dželila Mehanović, Zerina Mašetić, and Dino Kečo [35] focused on heart disease prediction using a majority-voting ensemble process. In this paper, they used the strategies of ANN, KNN and SVM to build a model to be used for forecasting cardiac disease. Using one of those algorithms, several studies are performed. Then, the ensemble's learning is incorporated, and the effects are compared. In both of these instances, by plurality voting, which is 61.16 percent for multi-class and 87.37 percent for binary grouping, the highest accuracy is achieved.

Javad Hassannataj Joloudari et al.[36] In 2020, a CHD disease diagnostic is proposed; important features are rated using a random tree model. The purpose of this research is to improve the accuracy of the diagnosis of CHD disease by selecting, according to their ranking, significant probabilistic features. They introduced an automated approach utilizing deep learning. The findings indicate that, relative to the other methods, the model of random trees is the better method for classification.

In this thesis, we focused on designing a new advanced machine learning method that would help a new concept that would vary in results and accuracy from other previous results and high performance to deal with medical data, especially CVD datasets.

3.2 OPTIMIZATION SYSTEM

Optimization refers to a method for finding a function's input parameters or arguments that result in the function's minimum or maximum performance. Continuous function optimization is the most popular method of optimization problems found in machine learning, where real-value numeric values, e.g. floating point values, are the input arguments to the function. A real-valued assessment of the input values is also the performance from the function [37]. To differentiate between functions that take different variables and are referred to as combinatorial optimization problems, we could refer to problems of this kind as continuous function optimization. There are several various forms of optimization algorithms that can be used, and maybe just as many ways to group and summarize, for continuous function optimization problems. One approach to grouping optimization algorithms is focused on the amount of accessible knowledge about the optimized goal function that can be used and harnessed by the optimization algorithm in turn. In general, the more knowledge known regarding the desired function, the better it is to refine the function if the information can be included in the search efficiently. Growing volumes of data generated in hospitals and medical services and the integration of various fields are contributing to a modern trend of precision and customized medicine for medical and healthcare study[38]. The trends offer a rare potential and good promise in medical and healthcare science to tackle diverse vital tasks. Such promise, however, depends heavily on whether we can identify helpful trends to define the target concerns, uncover insightful processes that underlie the noisy and scattered factual facts, and transform this insight into smart decision-making. Many attempts have previously been made across different methods, such as machine learning, computation, predictive analysis, mathematical simulation, biomedical computer science, etc [39]. This research would extend along the lines of methods of machine learning and optimization, and will create innovative models to address the problems of medicine or health care.

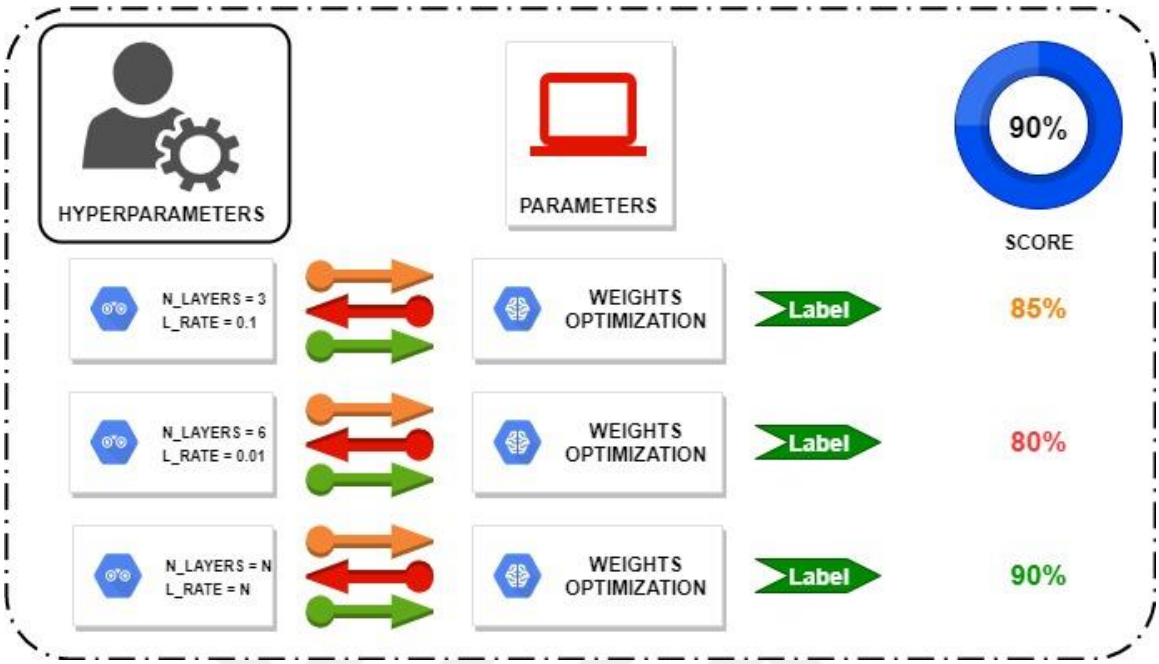

Figure 3. 1: Optimization Model Structure.

CHAPTER 4

4.1 METHODOLOGY

In this research, we work with advanced machine learning framework like advanced deep learning, super learner and SVM optimization to achieve a high-performance framework for the implementation of CVD 70000 data records for the diagnosis of CVD disease.

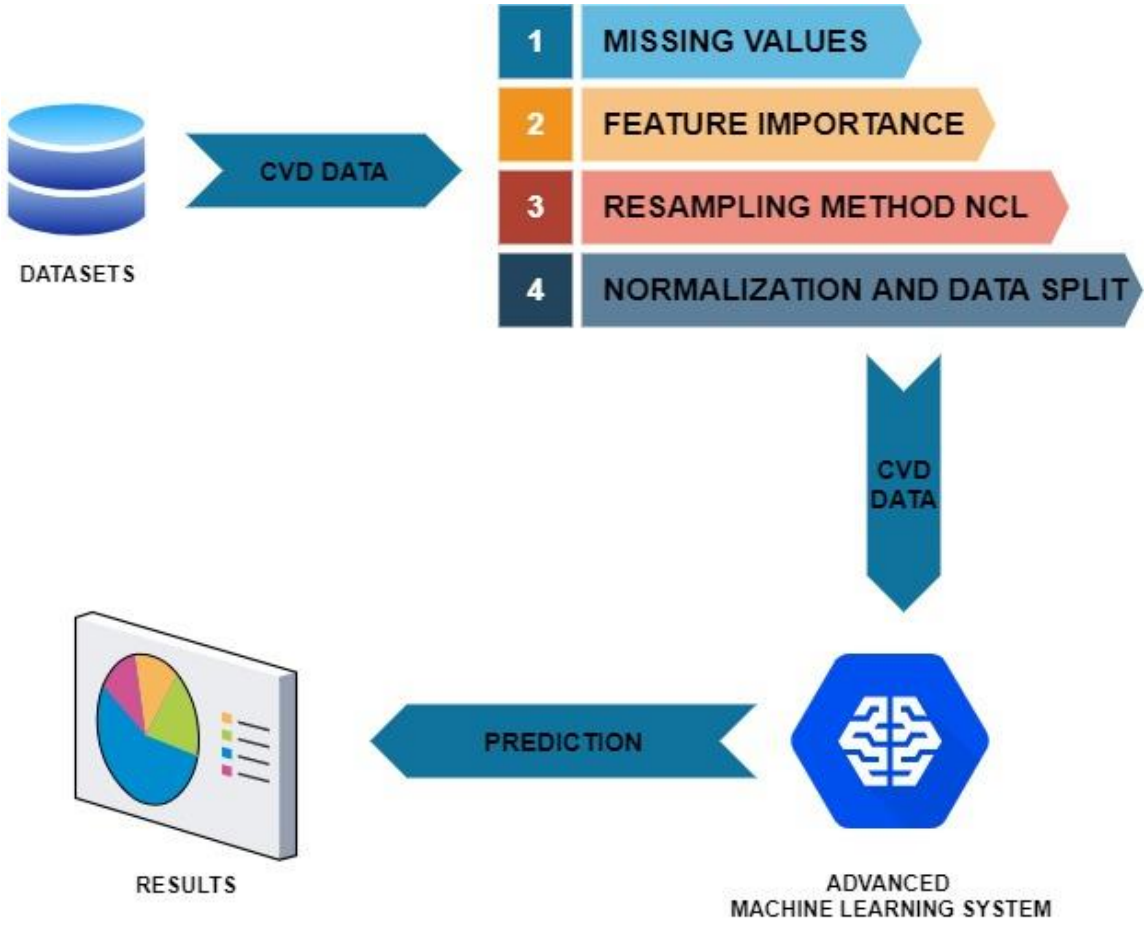

Figure 4. 1: Advanced Machine Learning Structure.

4.2 CVD DATASET DESCRIPTION

The CVD disease dataset contains 70,000 reference patient data (cardio) reports relating to the occurrence or absence of a 12-characteristic cardiac disorder, as seen in Table 1[40]. There are three forms of input characteristics: empirical (including factual information), examination (including findings of medical investigations) and subjective (including patient information). In Table 4.1 of the 70,000 records in this dataset, the objective variable is 'Cardio,' with 35,021 records in patients with cardio 0 and 34,979 records in patients with cardio 1.

Table 4. 1: The description of CVD dataset.

Attribute	**Kind**	**Disecription**
Id	Noumbers	Sequences
Age	Constant	Patient age in days
Gender	Discrete	1: female, 2: male
Height	Constant	Height of the patient in centimeters
Weight	Constant	Patient's weight in kilograms
Ap – hi	Constant	Blood pressure systolic
Ap – lo	Constant	Blood pressure in the diastole
Cholesterol	Distinct	1: average, 2: somewhat above normal, 3: significantly above normal
Gluc	Distinct	1 means average, 2 means somewhat above normal, and 3 means significantly above normal.
Smoke	Distinct	When a patient is a smoker or not
Alco	Distinct	Binary attribute of alcohol consumption
Active	Distinct	Binary attribute of physical exercise
Cardio	Distinct	CVD disorder (presence or absence)

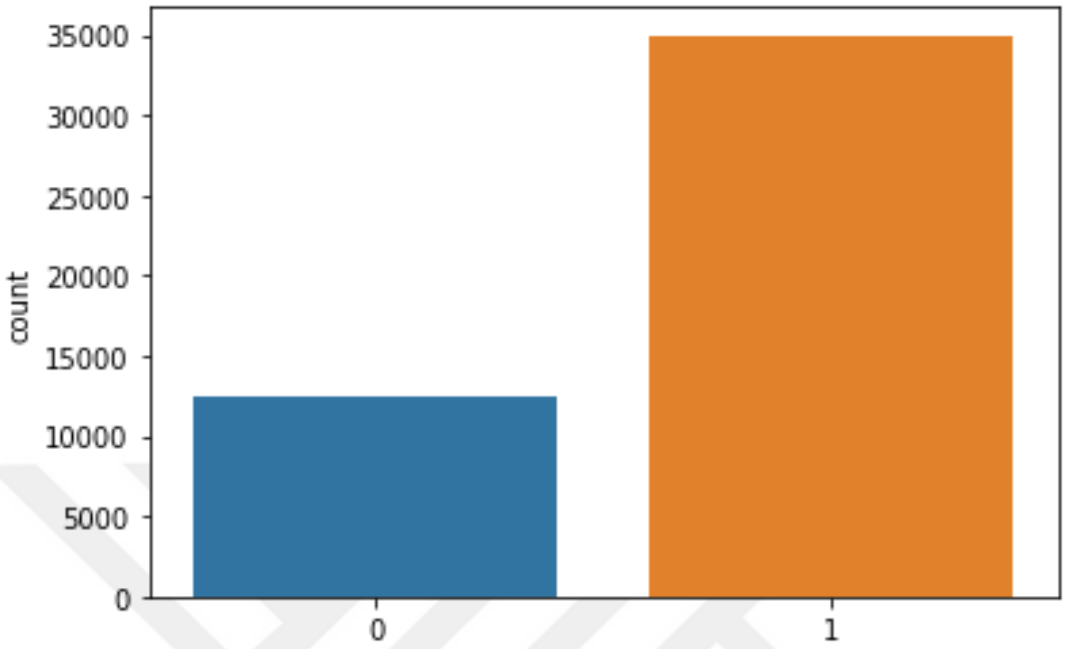

Figure 4. 2: The Distribution of Predictions Class.

4.3 DATA PREPROCESSING AND FEATURE OPTIMIZATION

Technology and collection of features, which takes into consideration the method of choosing the appropriate features from the CVD dataset and deleting features that do not allow a meaningful contribution to the goal variable prediction.[41]. It reduces the planning period and the quality of growth. We re-engineered redundant functionality with low over-the-counter CVD data priority that could impact our model's predictive outcomes [42]. For some data science tasks, especially some classification tasks, we used methods for the Random Forest ensemble and considered outstanding models. The explanations that this model is used may not need as much preprocessing as certain other approaches and may take as input categorical and numerical variables. Random forests and some other methods of boosting are basic, out-of-the-box models which we sometimes use early in the modeling phase to see how they operate with our data [43]. In comparison, recursive removal of features is used, which will be a greedy optimization approach that aims to build the best sub-set of features by continuously generating models and comparing the outputs of the model for each sub-set.

4.4 RESAMPLING USING HIGH PERFORMANCE MODEL

The distribution and class of CVD data sets are imbalanced and indistinguishable, since many natural processes seldom allow certain observations. In a community, for example, of limited diagnostic classes, CVD diseases may result in medical data. The models of machine learning will run into problems [44]. We combined class delivery with data reduction before the actual study, while we aimed at building an integrated machine learning framework. We used the latest form, neighbourhood cleaning rule (NCL), humble random and one-sided sorting techniques [45] in experiments with the CVD data collection. The Neighbourhood Cleaning Rule (NCL) will concentrate on cleaning up the data instead of utilizing the closest neighbourhood tool to condense the data and "editing" the dataset by removing samples that are not "full" with their neighbourhood. For each sample in the class to be sampled, the nearest neighbours are determined and, if the collection criteria is not satisfied, the sample is excluded [46]. Prior to re-sampling, the 70,000 records of 35,021 records were patients with Cardio 0 and 34,979 records were patients with Cardio 1, since 49331 records of 14352 records were patients with Cardio 0 and 34979 records were patients with Cardio 1.

4.5 OPTIMIZED SMOTE TECHNIQUE

A number of techniques are available for processing unbalanced data, including sample-based and algorithmic techniques, a mixture of sampling and algorithmic techniques, and function collection.[47]. The synthetic minority over-sampling method, in particular, is a form of synthesis re-sampling technique algorithm (SMOTE). Random over-sampling is used to choose the minority samples that will be replicated in a clustered manner [48]. In contrast to eliminating samples by under-sampling, some researchers choose to increase the number of samples because the latter can result in information loss. Random search is normally faster than grid search since it samples the search spectrum at random. It is claimed that random search is more efficient than grid search [49]. Figures 4.2 and 4.3 display the different distributions of expected class, with 49331 samples before over-sampling, including 14352 samples from patients with Cardio 0 and 34979 samples from patients with Cardio 1, and 34979 samples from patients with Cardio 0 and 1 after using the over-sampling process.

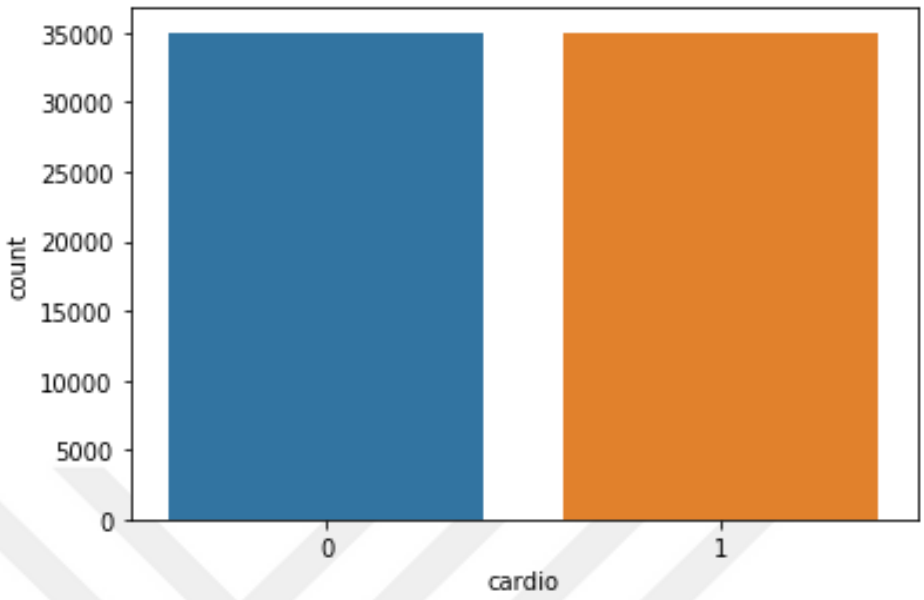

Figure 4. 3: The Distribution of Predicted Classes After Over-Sampling.

4.6 ADVANCED MACHINE LEARNING SYSTEM

This thesis explores a range of powerful methods that will get us closer to building an advanced model of difficult problems in machine learning [50]. We can create advanced deep learning models using the functional Keras API, share a layer across multiple inputs and use Keras models much like Python functions [51]. Much more important tasks would also be addressed, such as batch normalization, enhanced API features and hyper-parameter optimization.

4.6.1 Api Function Advanced

In the practical API, we can explicitly control tensors, and we use layers as functions that take tensors and return tensors [52]. We already learn from using the sequential to create models. The Functional API is a more flexible way to build templates than sequential ones [53]. It may accommodate models of non-linear topology, shared-layer models, and models of multi-input or multi-outputs.

4.6.2 Hyperparameter Optimization

So we need to scan the room of possible options instantly, consistently, principally. We ought to search for an architectural space to identify the best-performing ones empirically. That's what the world of automated hyper-parameter optimization is all about, a whole area of study and a significant one [54]. The algorithm using this validation performance background is the key to this process provided the numerous sets of hyper-parameters, learning rate, failure function, and epoch to choose the next collection of hyper-parameters to test. It is possible to do Bayesian optimization, genetic algorithms, basic random search and hyper-band tuner [55]. Using random sampling, hyper-band tuners were initially used to pick the right parameters to get a sense of good match, but smartly only persist after the initial results for the best performing candidates using previous results.

4.6.3 Batch Normalization

In order to improve our advanced machine learning platform and produce improved performance, we have used keras batch normalization. The batch size is a hyper-parameter that determines the amount of samples to operate through before the internal model parameters are told. Consider a batch with one or more specimens as a for-loop duplication and estimation [56]. The estimates are correlated at the end of the sample with the expected performance variables and an error is calculated. Via the batch normalization layer, Keras provides batch standardization support. To standardize them, the layer will transform inputs, indicating that they will have a mean of zero and a normal deviation of one [57]. During preparation, the layer can record statistics for each input variable and use them to standardize the results. Batch normalization may be used in certain models of deep learning neural networks, for certain points in a model. To standardize inputs before or after the activation function of the previous layer, the batch normalization layer can be used [58]. Batch normalization advantages enable the networks train quicker, allowing higher learning speeds, promoting weight initialization, simplifying the development of deeper networks, and delivering stronger overall performance.

4.7 SUPER LEARNER ADVANCED MODEL

The principle of a super machine learning model, or super learner, is appealing since it allows for the training and evaluation of a vast number of (ML) models on a single data set, enabling the algorithm to optimally integrate each of the individual models to provide better overall predictions.[59]. In medical science and spatial analysis, super learning has been used widely and has improved, although marginally, models of activity recognition from accelerometer. Instead, a new (ML) method identified at this time can provide a key to the problem of method selection [60]. Super Learning requires a collection of candidate learners (other ML methods), incorporates them in a data set, and chooses a fitting mix of learners based on the resulting cross-validated method. In order to do as good as all of the learner inputs, the super learner model (SL) attempts to find the strongest candidate learner mix. The ML-Ensemble, using the same general trend for classification issues, is also really simple to use [61]. In this scenario, we can use the super learner as the meta-model to categorize a list of classifier models and a logistic regression model with the ML-Ensemble library. The full effects of fitting and testing a super learner model for a test classification issue with the (mlens) library are described below in the section on Results base-model consistency, and eventually super-learner outcomes on the holdout dataset. Considering the stochastic nature of the CVD dataset and learning algorithms, the real findings will vary.

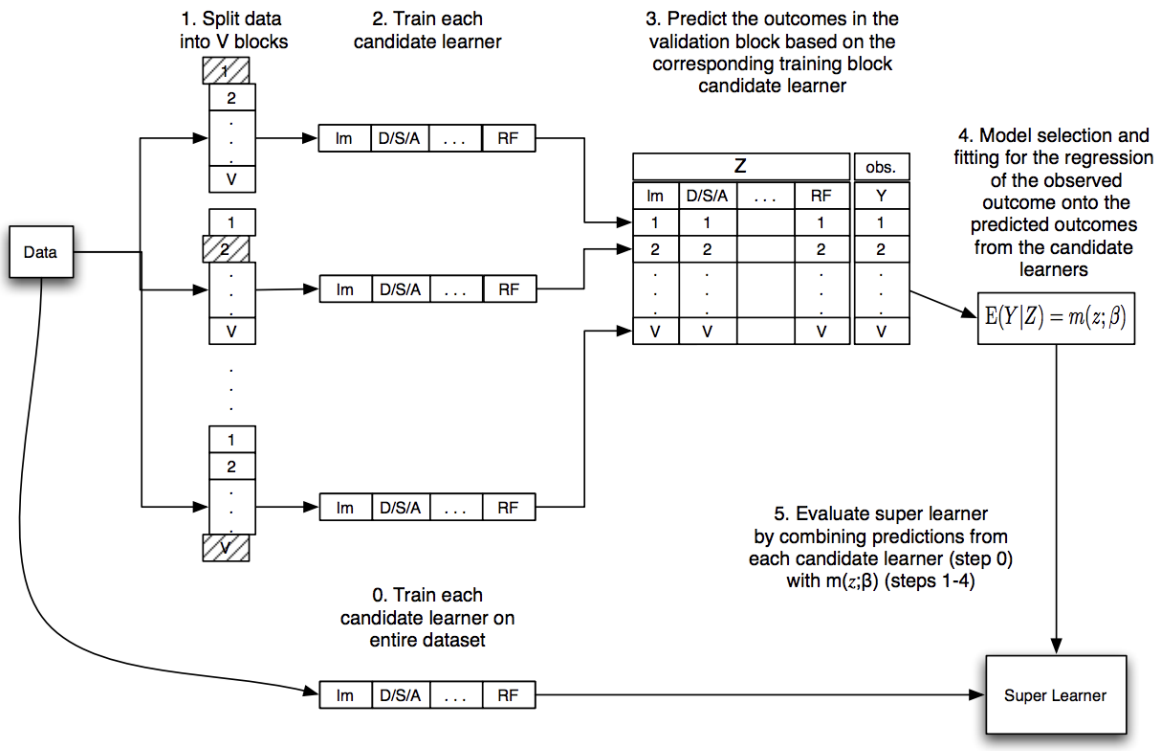

Figure 4. 4: Super Learner Advanced Model Structure [62]**.**

4.8 SVM HYPERPARAMETER OPTIMIZATION

One of the top methods of classification problems in medicine has lately been known to be SVM. It has many attractive features, such as a clear mathematical base, few tuning parameters, quick classification and high potential for generalization. Although SVM does not consider adequate in the default parameter to have advanced and high-performance CVD diagnostic issue machine learning model so that SVM needs to be optimized, including hyper-parameter tuning [63]. For quest space discovery, we used the most powerful technique through a uniformly spaced grid called Grid Search (GS) hyper-parameter tuning, including three main sides such as the kernel, regularization, and gamma.

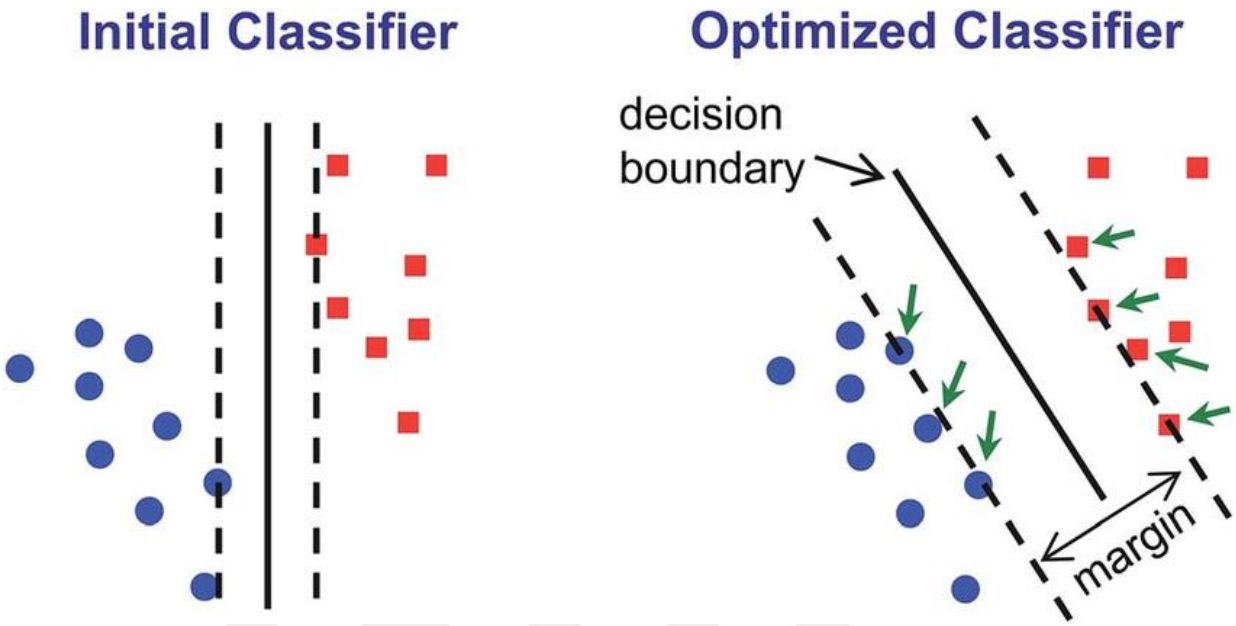

Figure 4. 5: SVM Optimization Model Structure [64].

The key role of the kernel is to transform the data input to the required type in the specified dataset [65]. There are numerous forms of function, such as linear, polynomial and radial base functions (RBF). Polynomials and RBF are useful for non-linear hyper-planes. The regularization feature is used in the Scikit-learn C python module to retain regularization. C is the parameter of the penalty here, reflecting the misclassification or error of the term. When optimizing the SVM, the term misclassification or error means how much error it may bear [66]. Gamma would approximately balance the training data set with a lower Gamma value, whereas a higher gamma value will exactly suit the training data set, which causes over fitting. In other terms, we can infer that a low gamma value just includes neighboring points in the estimation of the separation line, while the gamma value takes into consideration all the data points when measuring the separation line.

4.9 AUTO SKLEARN MODEL

This thesis explores the popular generic ML toolbox, Scikit-Learn, promoted by the name of Auto-Sklearn (Auto-ML). Auto-ML uses the Scikit-Learn toolbox's numerous ML classifiers (14 in total) and pre-processing measures (14 function processing strategies and four data pre-processing methods) to simplify the creation of an AI model.[67]. Logistic regressions, SVM, RF, Boosting and neural networks are interested in this. Auto-ML defines Auto-ML as the process of automatically making test set predictions if a given computational budget is provided (Without interference by humans). The measurement budget is the amount of time or memory needed for the problem to be solved by the machine [68]. To identify the right AI structures and parameters, Auto-ML mixes standard machine learning methods with a Bayesian optimization model.

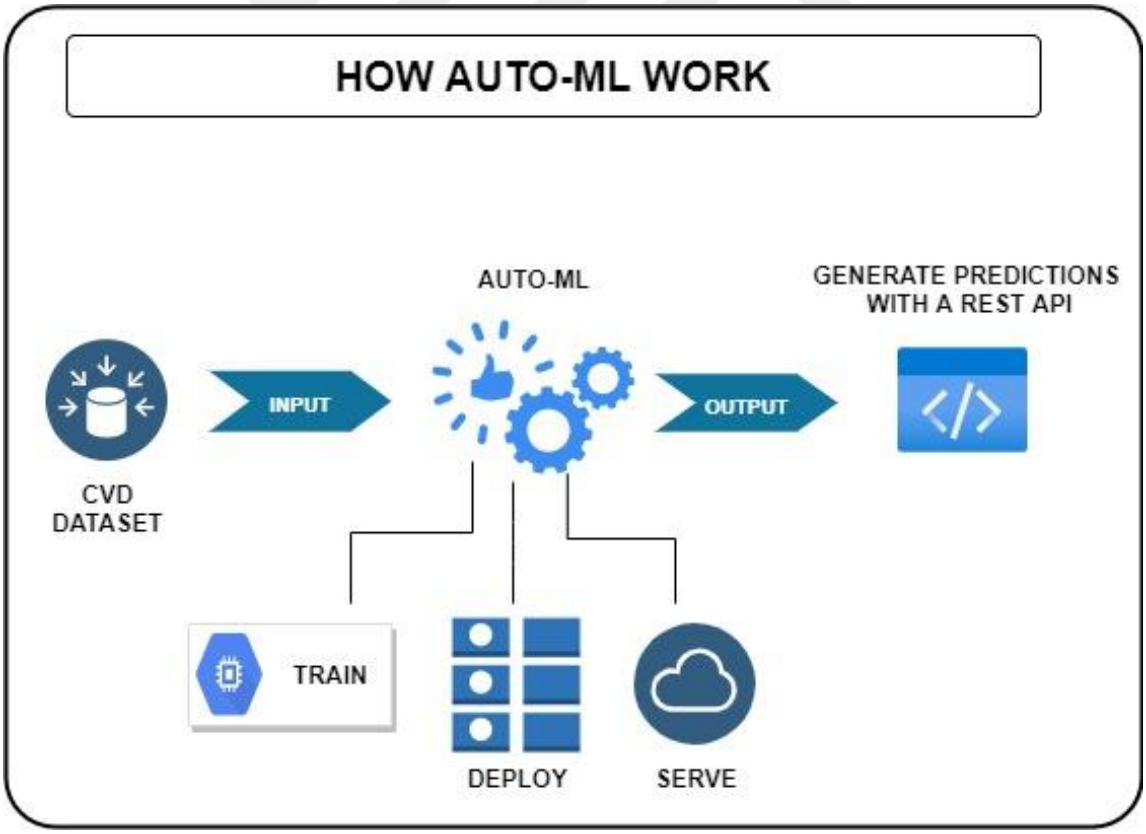

Figure 4. 6: Auto-ML Advanced Model Structure.

4.10 ADVANCED XGBOOST

XGBoost has been one of the most used methods in ML. It consists of a collection of decision trees, where each current tree is focused on the evaluation of the previous one. Every machine learning model has hyper-parameters, parameters that in the training process are not optimized [69]. This means that we have to try a variety of combinations by hand to find optimum values. We applied the hand optimization model in this analysis, including the simple and highly efficient Grid Search, to the two most common Grid Search and Random Search techniques. Grid Quest is a simple way to define ideal hyper-parameters by determining some mixture of hyper-parameters. The number of iterations is the product of the number [70] of a hyper-parameter. When defining the XGBoost classifier, we used a range of parameters, such as learning intensity, max-depth, n-estimators, gamma, and alpha.

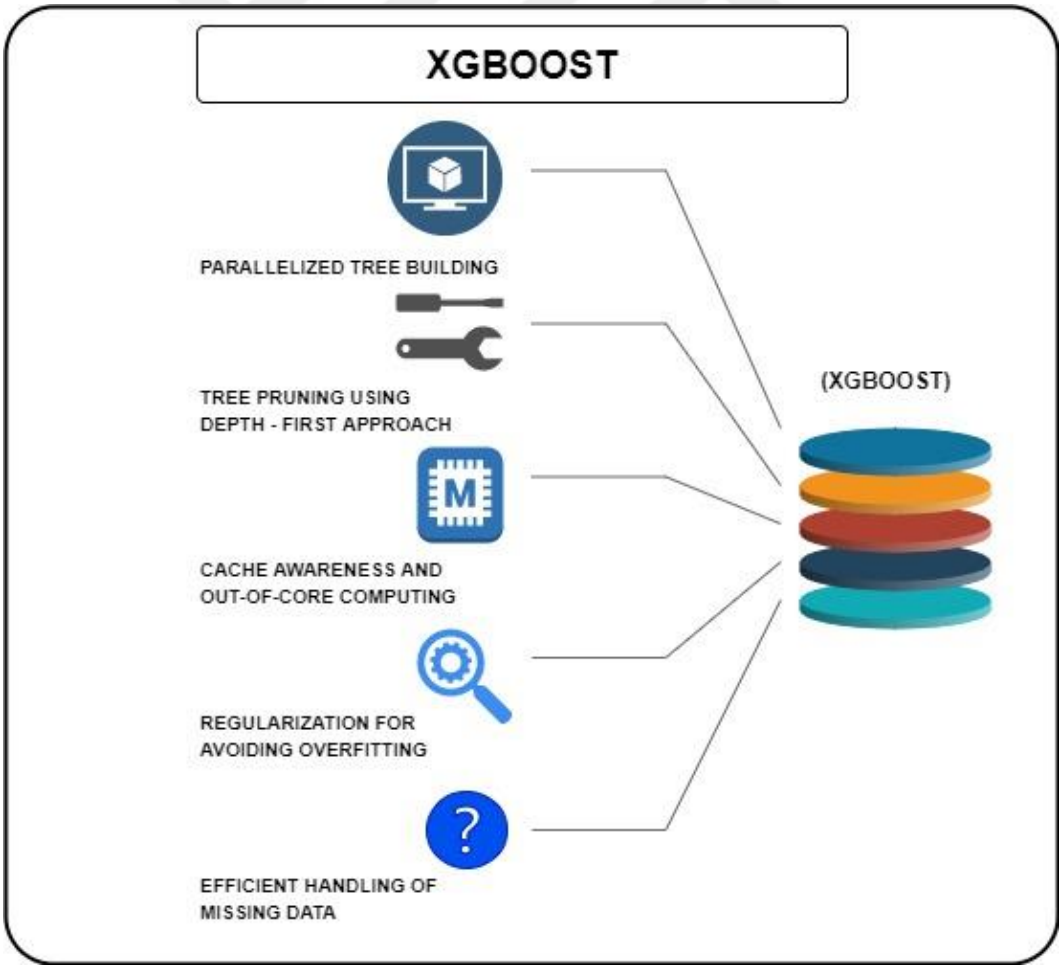

Figure 4. 7: XGBoost Advanced Model Structure.

4.11 ADVANCED STACKING MODEL

When we first discussed stacking theory, we stated that the aim is to practice with a large number of weak learners and then combine them by training a meta-model to return predictions equivalent to the number of predictions provided by the stack of weak learners.[71]. There are two major differences in stacking, bagging, and boosting. The algorithms that train heterogeneous poor learners (learning algorithms with different parameters) are more general when it comes to cluster construction.

4.12 ADVANCED ENSEMBLE BAGGING MODEL

Bagging stands for sorting by bootstrap. One of the oldest, most interesting and most powerful algorithms in an ensemble are bagging. The central theory of bagging is using bootstrapped replicas of the initial dataset and using them to train different classifiers [72]. A simple majority vote is commonly used in a classification system to aggregate all base classifiers into a single model to provide the final performance of the ensemble model, taking the total of all regression model predictions. A basic definition of such a technique is provided by the Random Forest algorithm. Bagging decreases the high variance of a model, thus reducing the generalization error [73].

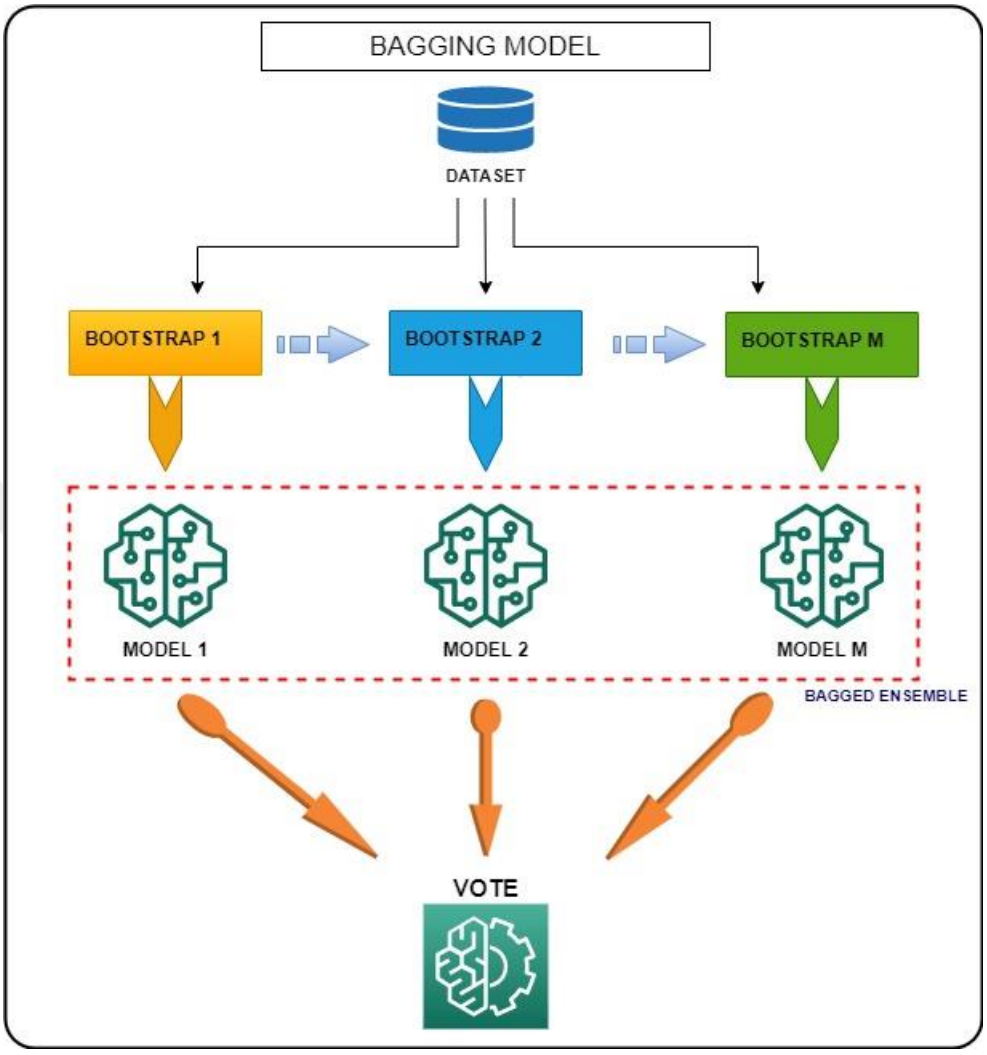

Figure 4. 8: Bagging Model Structure.

Bagging, especially when you have very little experience, is a very good technique. Through aggregating the scores over many experiments with the use of bootstrapped samples, we are able to get an estimate.

4.12 OPTIMIZED ADABOOST

AdaBoost (Adaptive Boosting) is a very common boosting technique that attempts to construct one strong classifier by merging several weak classifiers. A single classifier might not be able to correctly predict an object's class, but if we group several poor classifiers with each gradually learning from the wrongly categorized artifacts of the others, we can create one such efficient model. One of your simple classifiers might be the classifier listed here, from Decision Trees (often the default) to Logistic Regression, etc [74]. Now we could inquire, what is a "bad" classifier? A bad classifier is one that does better than random guessing, but also does badly in designating entity classes. A bad classifier, for example, will predict that anyone above the age of 40 cannot run a marathon, but those who fall below that age can. You might get over 60% precision now, but you'd also misclassify a ton of data points. AdaBoost should be implemented on top of any classifier rather than being a model of itself, to benefit from its limitations and propose a more precise model. For this function, it is normally named the "best out-of-the-box classifier". Let's try to explain how Judgment Stumps operates with AdaBoost. In a Random Woodland, Decision Stumps are like trees, just not "completely developed." They have one node and two leaves. Rather than plants, AdaBoost uses a landscape of such stumps. Stumps alone are not a successful decision-making tool [75]. To estimate the goal value, a full-grown tree incorporates the decisions of all variables. In the other side, a stump should only use one element for a judgment to be made. Let's try to explain step-by-step the behind-the-scenes of the AdaBoost algorithm by looking at several factors to decide whether or not an individual is "fit" (in good health). AdaBoost has a number of benefits, mainly, unlike algorithms such as SVM, it is easier to use with less need for manipulating parameters. As a perk, with SVM, you can use AdaBoost as well. AdaBoost is technically not vulnerable to over-fitting, but there is no clear evidence for this. It is possible to use AdaBoost to increase the accuracy of your poor classifiers, rendering it versatile [76]. It has now been expanded past binary classification, and usage cases have now been established in text and picture classification.

CHAPTER 5

5.1 RESULTS BEFORE RESAMPLING

5.1.1Advanced Deep Learning

Using the strongest python programming framework, we applied our sophisticated machine learning method in multiple strands. Next, we used data preprocessing and feature significance utilizing techniques for the Random Forest ensemble that brought us the best 11 attributes we worked with. Secondly, we introduced an advanced deep learning model with CVD 70000 data records with 35,021 records for Cardio 0 patients, including hyper-parameter optimization, and 34,979 records for Cardio 1 patients with 70 percent for preparation and 30 percent for test scale. In order to choose the next range of hyper-parameters to evaluate using the hyper-band model, this optimization method involves learning rate, loss function, and epoch. In Figure 5.1, the outcomes of this step can be seen.

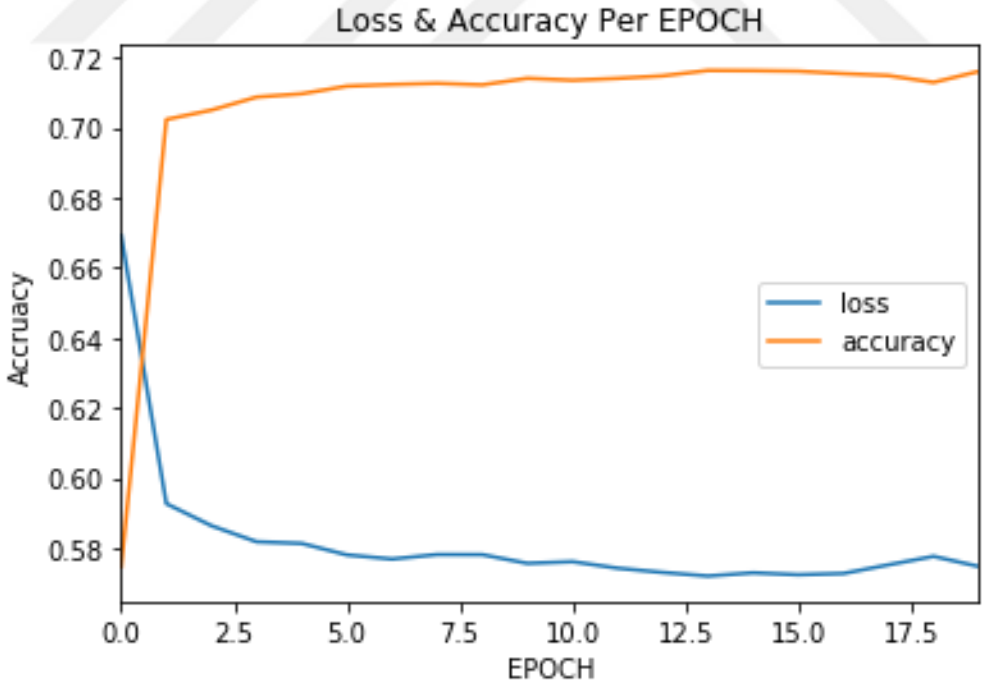

Figure 5. 1: Showed The Optimization Results Using Hyper-band Model.

Figure 5.1 showed the better accuracy of the optimization stage, which was around 72 percent after we introduced advanced and batch normalization of the API feature, which improved the efficiency and accuracy of our advanced method. Figure 5.2 indicates the results of batch normalization.

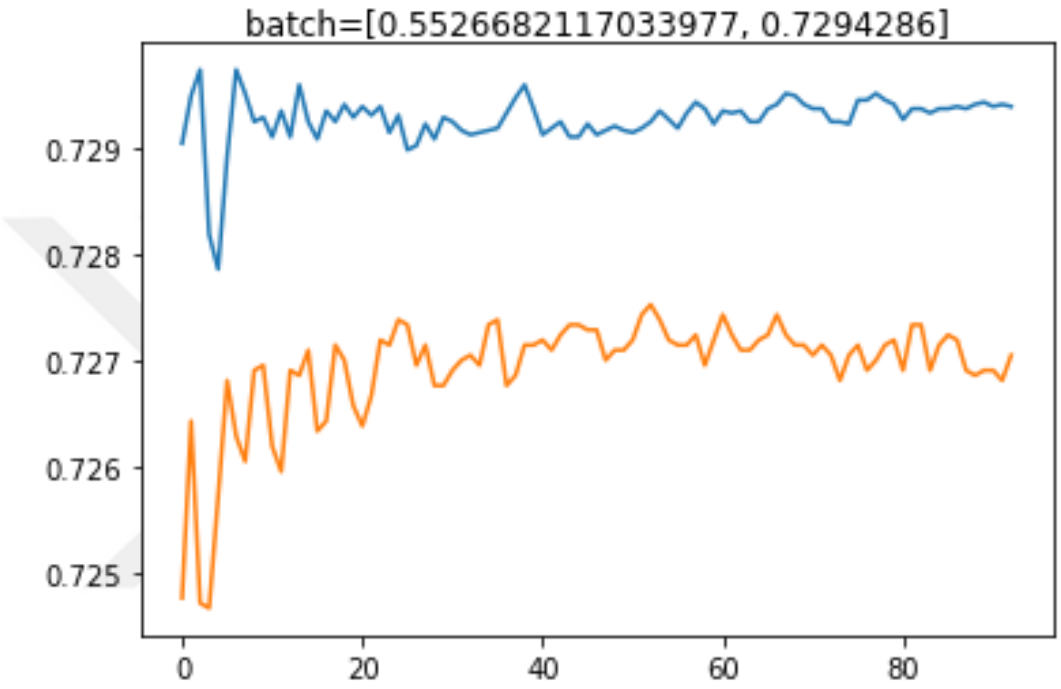

Figure 5. 2: Showed The Batch Normalization Results Using.

We see the advantages and the effect of utilizing batch normalization and advanced API feature on the model accuracy, which increased to 73%, in the figure 5.2. We provided the training and validation precision of our advanced model in figures 5.3 and 5.4 with the best optimized epoch and training failure and validation loss.

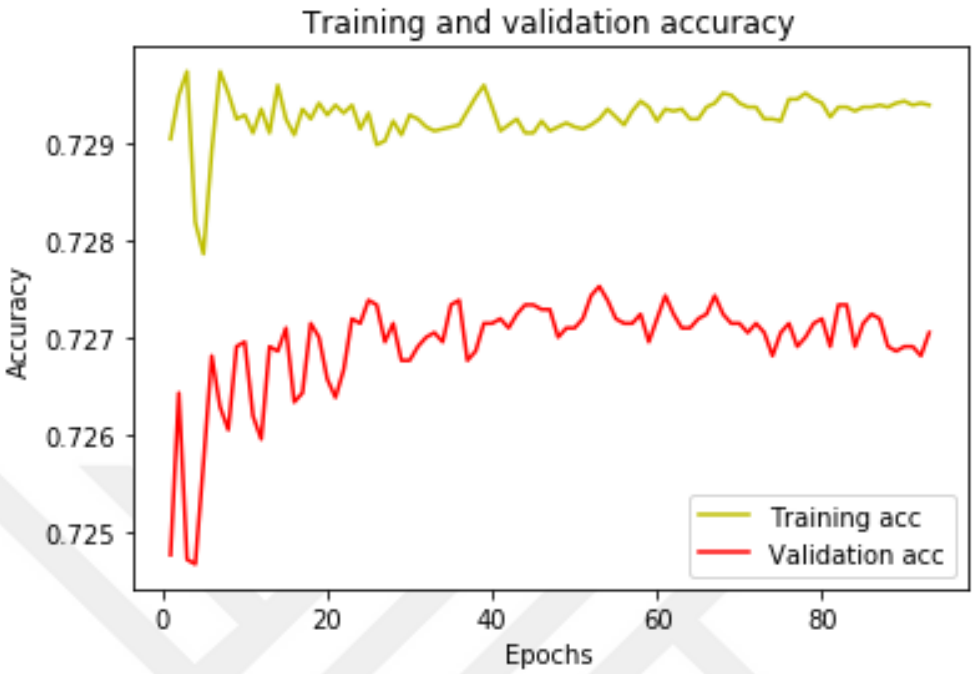

Figure 5. 3: Showed The Training and Validation Accuracy.

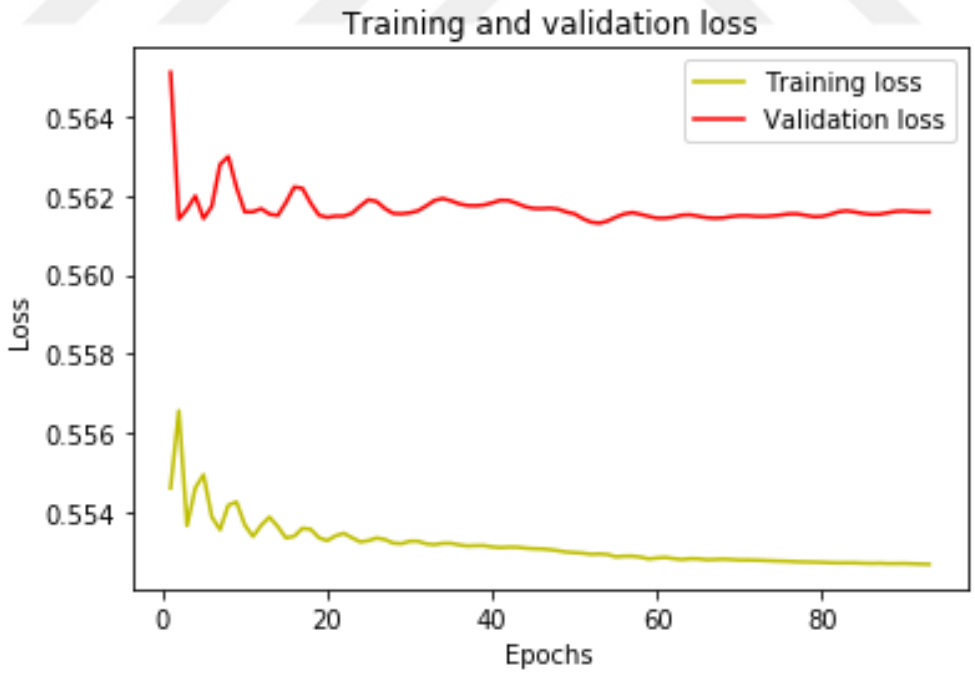

Figure 5. 4: Showed The Training and Validation Loss.

5.1.2 Advanced Super Learner

We introduced the super learner with a collection of classifier models and a logistic regression as the meta-model to classify with the ML-Ensemble library. With 70000 CVD data records, the outcomes of this model obtained 76 percent precision like train like 49000 and test 21000 with the classification report provided in table 5.1.

Table 5. 1: The results of super learner model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.75	0.76	0.76	10509
Cardio 1	0.76	0.75	0.75	10491
Avg.	0.76	0.76	0.76	21000

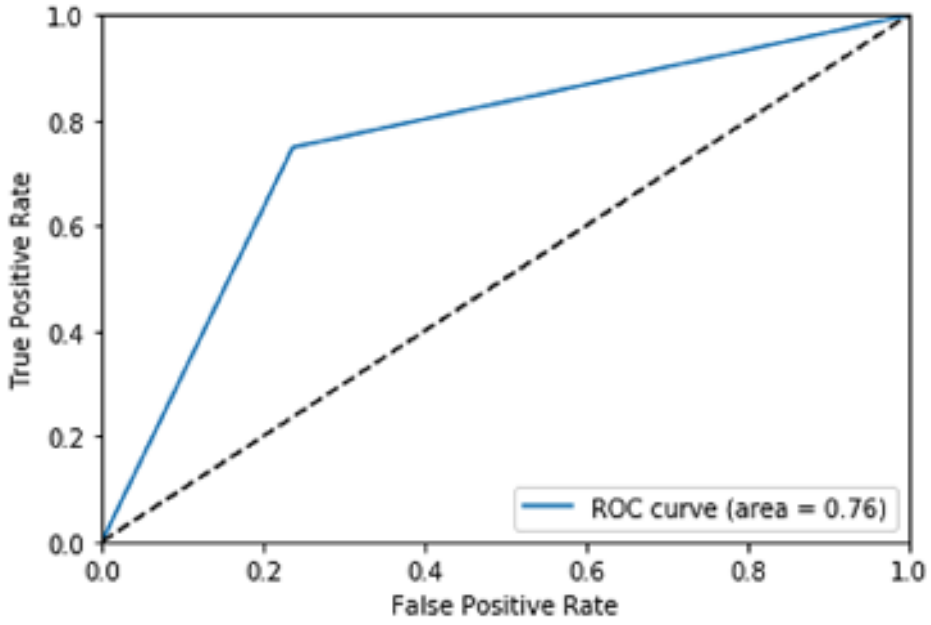

Figure 5. 5: Showed The ROC Curve of Super Learner Model.

In the figure 5.5 we presented ROC curve area of super learner which got 76% accuracy using 70000 CVD data records.

5.1.3 Auto-Ml Model

We implemented our smart high-performance machine learning approach in various stages utilizing the best python programming framework. Next, for the Ensemble Random Forest ensemble, we used data pre-processing and function importance using approaches that got us the best 11 attributes we worked with. Secondly, with 35,021 Cardio 0 patient records and 34,979 Cardio 1 patient records, we presented the Auto-Sklearn (AutoML) model with CVD 70000 data records, with 70 percent for testing and 30 percent for test scale. In Table 5.2, the outcomes of this step can be seen.

Table 5. 2: The results of AutoML model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.70	0.79	0.74	10598
Cardio 1	0.76	0.66	0.71	10402
Avg.	0.73	0.73	0.73	21000

5.1.4 Optimized SVM

We implemented SVM hyper-parameter tuning in order to achieve a high performance model in this segment. Via a uniformly spaced grid called Grid Quest (GS) including kernel tuning, we used the most powerful search space exploration technique. We received 71 percent improved precision with 70000 of CVD data records with the strongest parameter set on creation including ('C': 100,' kernel':' linear) with 70000 of CVD data records. The results of the classification report of the configured SVM model are summarized in Table 5.3.

Table 5. 3: The results of optimized svm model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.68	0.80	0.74	10539
Cardio 1	0.75	0.62	0.68	10461
Avg.	0.71	0.71	0.71	21000

5.1.5 Advanced XGBoost

In order to classify with the ML-Ensemble library, we used the advanced XGBoost model. With 70000 CVD data records with best parameters used ('learning rate': 0.1,' max depth': 4,' n estimators': 100,' cross-validation ': 10 folds), the effects of this model achieved 73 percent accuracy including train like 49000 and test 21000 with classification report provided in Table 5.4.

Table 5. 4: The results of advanced XGBoost model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.71	0.79	0.74	10598
Cardio 1	0.75	0.67	0.71	10402
Avg.	0.73	0.73	0.73	21000

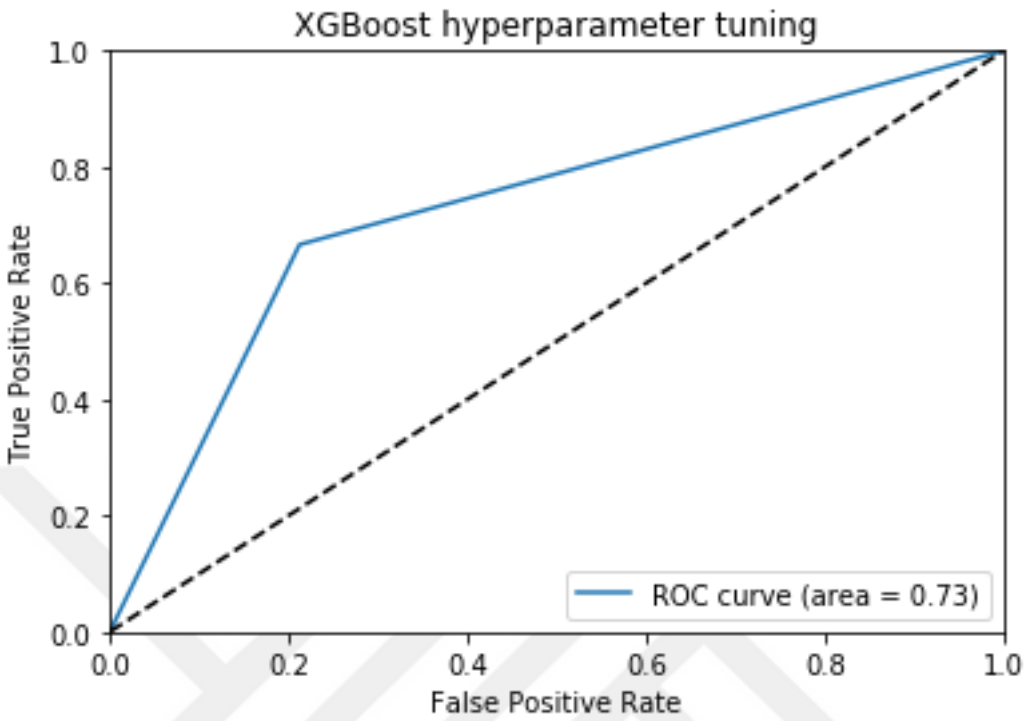

Figure 5. 6: Showed The ROC Curve of Advanced XGBoost Model.

5.1.6 Advanced Ensemble Bagging Model

We used advanced ensemble bagging model in order to attain advanced high performance model in this segment. With 70000 CVD data reports, with the better bagging model, we got 73 percent greater precision. In table 5.5, we presented the findings of the advanced ensemble bagging method with model form and accuracy.

Table 5. 5: The results of advanced ensemble bagging model.

Model Type	**Model Accuracy**
Normal Random Forest Classifier	0.6759 (+/- 0.0053)
Bagging Random Forest Classifier	0.7117 (+/- 0.0013)
Normal Extra Trees Classifier	0.6669 (+/- 0.0023)
Bagging Extra Trees Classifier	0.6977 (+/- 0.0018)
Normal K-Neighbors Classifier	0.6723 (+/- 0.0016)
Bagging K-Neighbors Classifier	0.7031 (+/- 0.0039)
Normal AdaBoost Classifier	0.7259 (+/- 0.0024)
Bagging AdaBoost Classifier	0.7261 (+/- 0.0025

5.1.7 Advanced AdaBoost Model

We implemented AdaBoost hyper-parameter tuning in order to achieve a high performance model in this segment. Via a uniformly spaced grid called Grid Search (GS) including the most powerful search space exploration technique. We received 73 percent improved precision with 70000 of CVD data records with the strongest parameter set on creation including ('learning_rate': 0.1, 'n_estimators': 1000) with 70000 of CVD data records. The results of the classification report of the configured advanced AdaBoost model are summarized in Table 5.6.

Table 5. 6: The results of advanced AdaBoost model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.70	0.80	0.75	10598
Cardio 1	0.76	0.65	0.70	10402
Avg.	0.73	0.73	0.73	21000

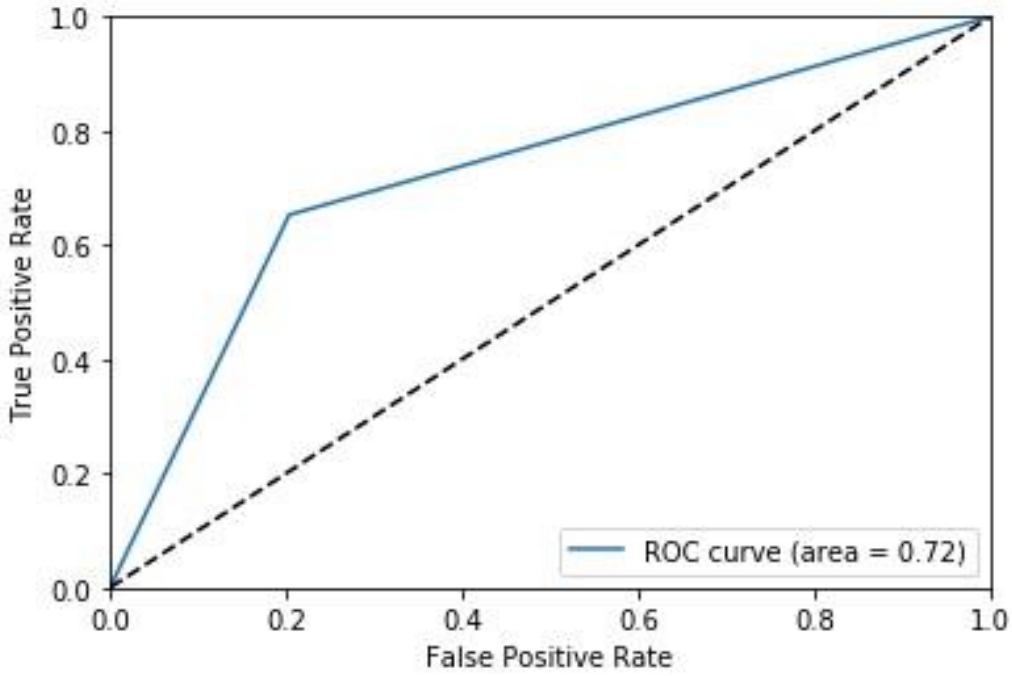

Figure 5. 7: Showed The ROC Curve of Advanced AdaBoost Model.

CHAPTER 6

6.1 RESULTS AFTER RESAMPLING

We also applied the neighborhood cleaning rule (NCL) that takes into account the high performance CVD dataset process. We have used a split train test to split our dataset into a train and test with 30 percent of the CVD dataset as a test scale and 70 percent as a testing size. In addition, we used preprocessing for our dataset, such as preprocessing of attribute relevance and missing values. After re-sampling process, the distribution of predictable class (Cardio) was 49331 data records consisting of 2 classes 0 with 14352 and 1 with 34979.

6.1.1 Advanced Deep Learning

After re-sampling, the results of our advanced deep learning model that we applied with CVD data improved to 79 percent accuracy with hyper-band optimization and we received 80 percent as the best accuracy of our advanced machine learning framework after API working with batch normalization. The effects of advanced deep learning are seen in table 6.1 below as a classification report after the optimization phase.

Table 6. 1: The results of advanced deep learning model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.64	0.60	0.62	4254
Cardio 1	0.84	0.87	0.85	10546
Avg.	0.79	0.79	0.79	14800

In addition we display the accuracy of optimization process in the figure 6.1.

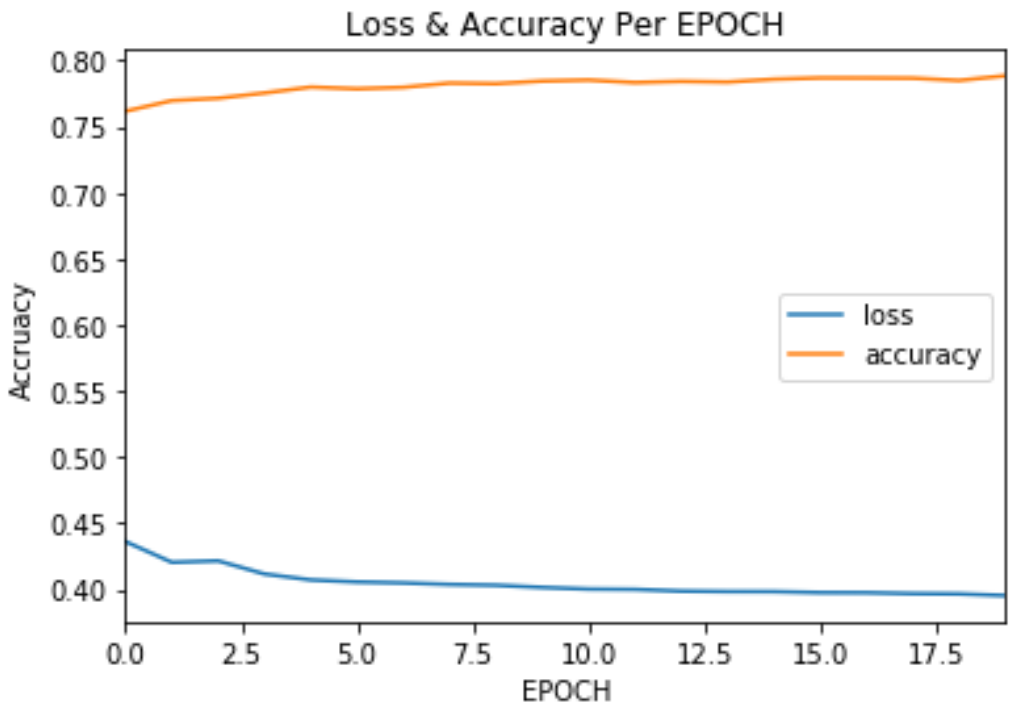

Figure 6. 1: Showed The Optimization Accuracy Using Hyper-band Model with Resampled CVD Data.

The figure 6.1 shows the optimization phase better accuracy, which was around 79 percent after we introduced advanced API feature and batch normalization with re-sampled CVD data that improved our advanced efficiency and accuracy of the method. Figure 6.2 indicates the results of batch normalization.

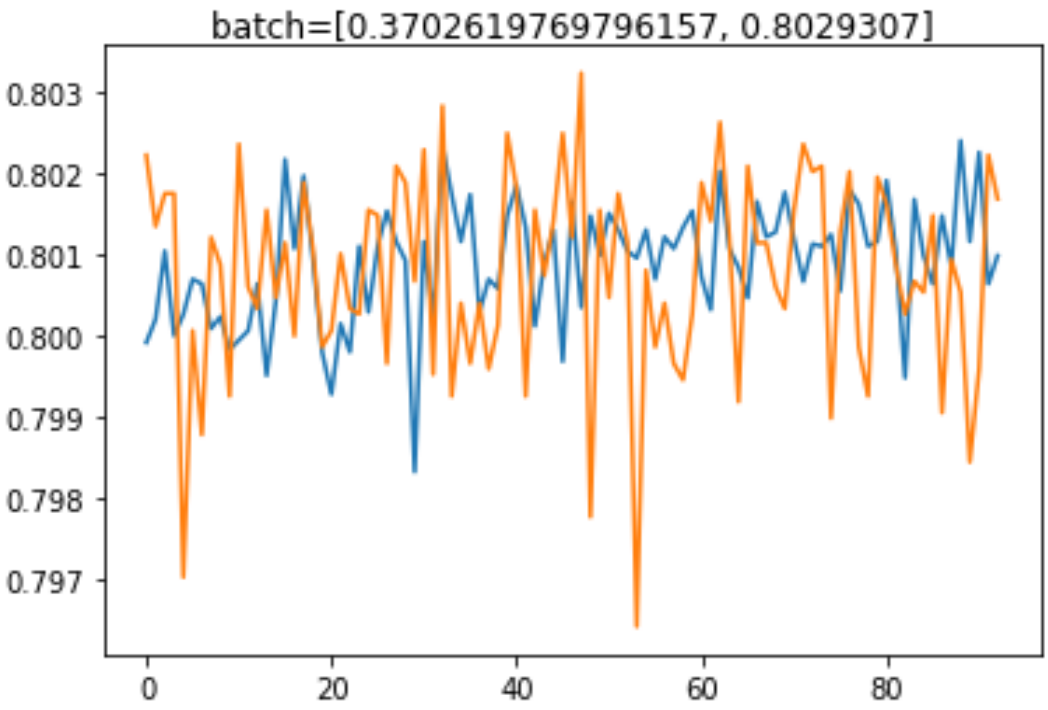

Figure 6. 2: Showed The Batch Normalization Results with Resampled CVD Data.

The precision of advanced deep learning with the best epoch, learning rate batch size, and best number of layers in figure 6.3 with batch normalization and API feature.

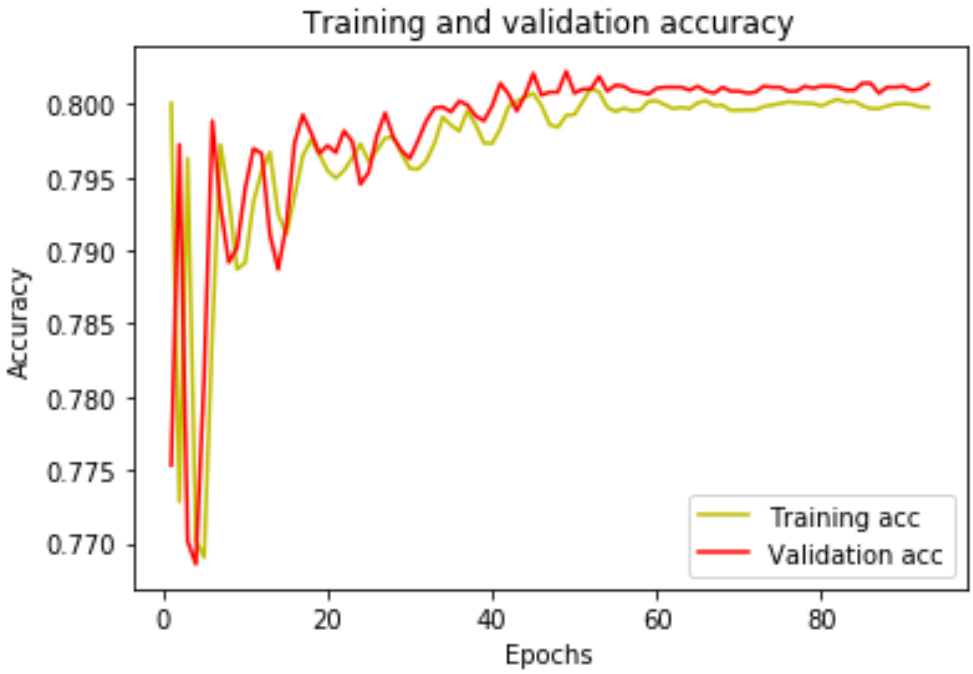

Figure 6. 3: Showed The Training and Validation Accuracy of Advanced Deep Learning Model with Resampled CVD Data.

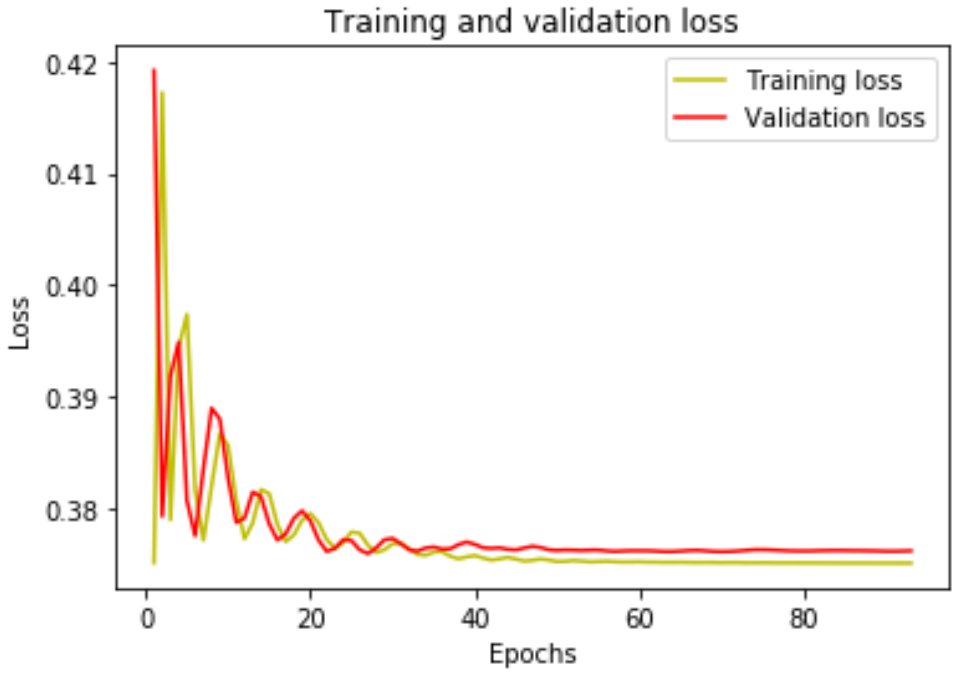

Figure 6. 4: Showed The Training and Validation Loss of Advanced Deep Learning Model with Resampled CVD Data.

6.1.2 Advanced Super Learner Model

With 49331 CVD data reports, the outcomes of this model obtained 79 percent precision like train like 34531 and test 14800 with this model's classification study with 49331 CVD data provided in table 6.2.

Table 6. 2: The results of super learner model with resampled CVD data.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.78	0.80	0.79	7347
Cardio 1	0.80	0.77	0.79	7453
Avg.	0.79	0.79	0.79	14800

In the figure 6.5, we displayed a super learner's ROC curve region with 79% precision using 49331 CVD data records.

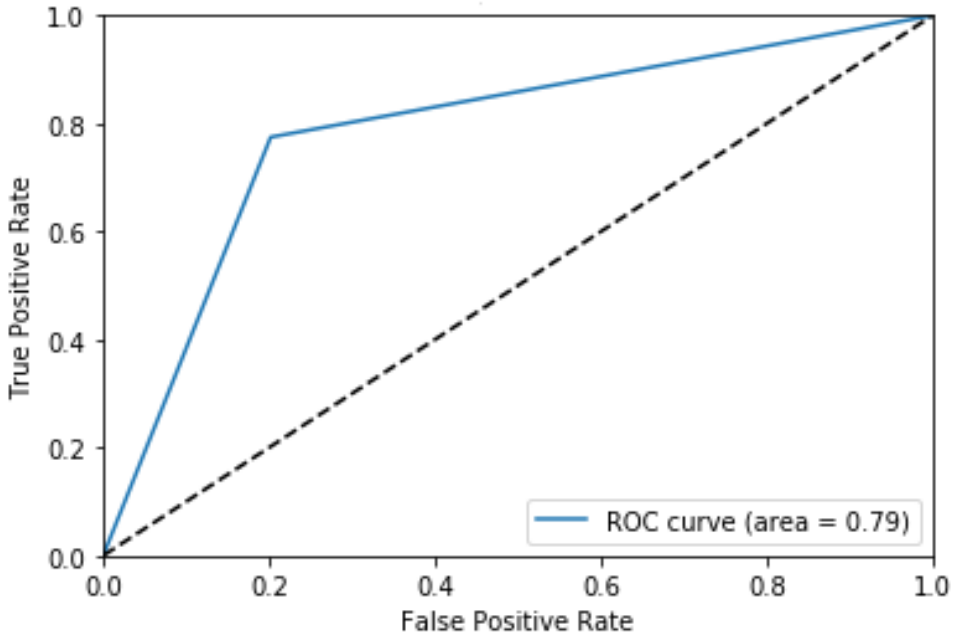

Figure 6. 5: Showed The ROC Curve of Super Learner Model.

6.1.3 Optimized SVM

We implemented SVM hyper-parameter tuning in order to achieve a high performance model in this segment. Via a uniformly spaced grid called Grid Quest (GS) including kernel tuning, we used the most powerful search space exploration technique. We received 80 percent improved precision with 49331 of CVD data records with the strongest parameter set on creation including ('C': 100,' gamma': 0.001,' kernel':' rbf') with 49331 of CVD data records. The results of the classification report of the configured SVM model are summarized in Table 6.3.

Table 6. 3: The results of optimized SVM model with resampled CVD data.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.63	0.75	0.69	4254
Cardio 1	0.89	0.82	0.86	10546
Avg.	0.82	0.80	0.81	14800

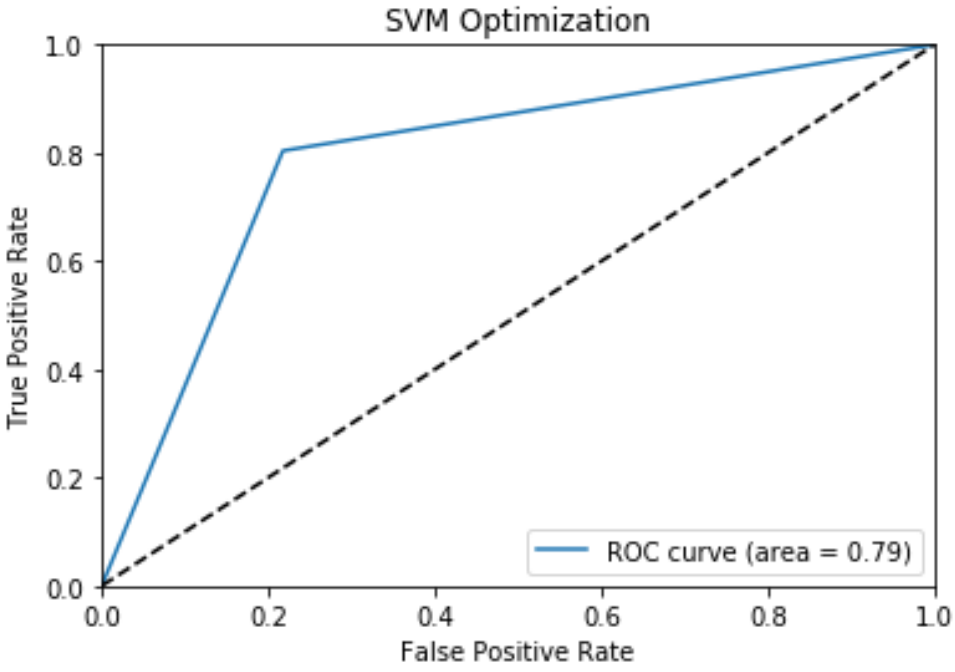

Figure 6. 6: Showed The ROC Curve of Optimized SVM Model with Resampled CVD Data.

In the above figure, we showed an optimized SVM model ROC curve region that achieved 79 percent accuracy using 49331 CVD data records.

6.1.4 Auto-Ml Model

After re-sampling, the results of the Auto-ML model we applied with CVD data improved to 83 percent accuracy with hyper-band optimization and we received 80 percent as the best accuracy of our advanced machine learning framework after API working with batch normalization. The effects of advanced Auto-ML are seen in table 6.4 below as a classification report after the optimization phase.

Table 6. 4: The results of AutoML model before SMOTE.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.69	0.76	0.72	4357
Cardio 1	0.90	0.86	0.88	10443
Avg.	0.83	0.83	0.83	14800

Table 6. 5: The results of AutoML model after SMOTE.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.83	0.92	0.88	10468
Cardio 1	0.91	0.81	0.86	10520
Avg.	0.87	0.87	0.87	20988

In addition we display the ROC curve of AutoML model in the figure 6.7.

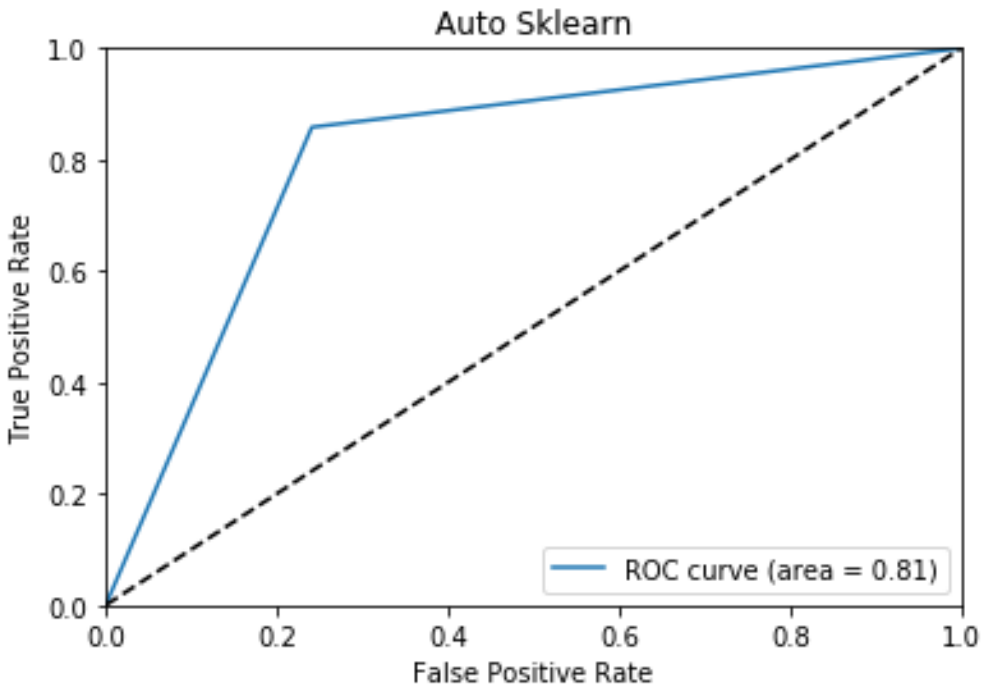

Figure 6. 7: Showed The ROC Curve of AutoML Model Before SMOTE.

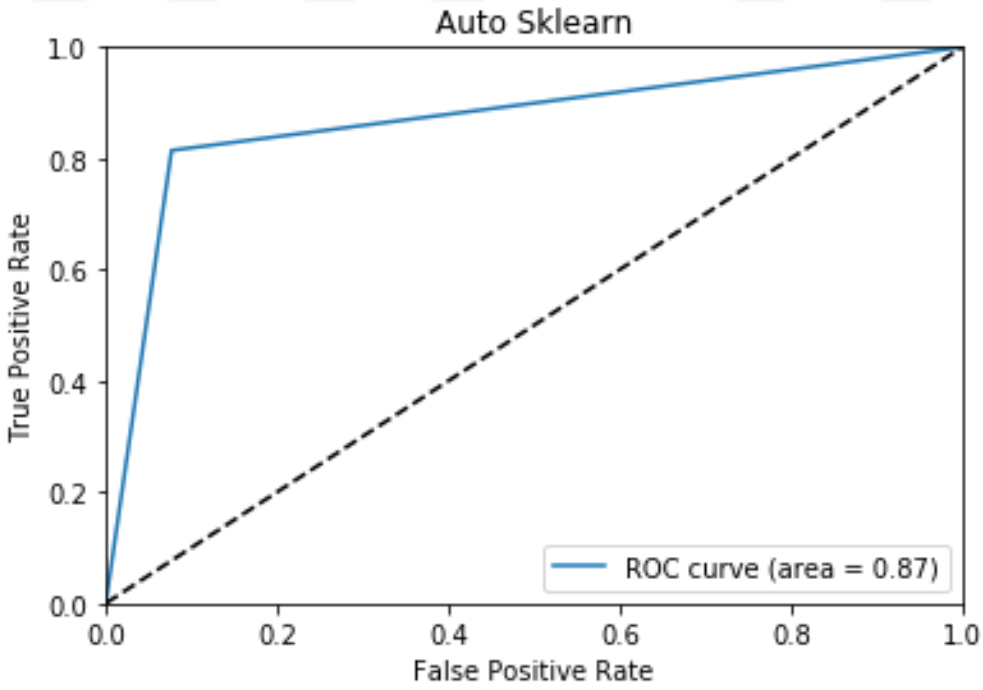

Figure 6. 8: Showed The ROC Curve of AutoML Model After SMOTE.

6.1.5 Advanced XGBoost Model

The results of this model obtained 81 percent accuracy with 49331 CVD data reports, such as train like 34531 and test 14800 with ('learning rate': 0.01,' max depth': 10, number of trees': 500,' cross-validation': 10 Folds) classification report of this model with 49331 CVD data given in table 6.5.

Table 6. 6: The results of advanced XGBoost model with resampled CVD data.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.65	0.75	0.70	4254
Cardio 1	0.89	0.84	0.87	10546
Avg.	0.82	0.81	0.82	14800

Table 6. 7: The results of advanced XGBoost model after SMOTE CVD data.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.81	0.94	0.87	10528
Cardio 1	0.93	0.78	0.85	10460
Avg.	0.87	0.86	0.86	20988

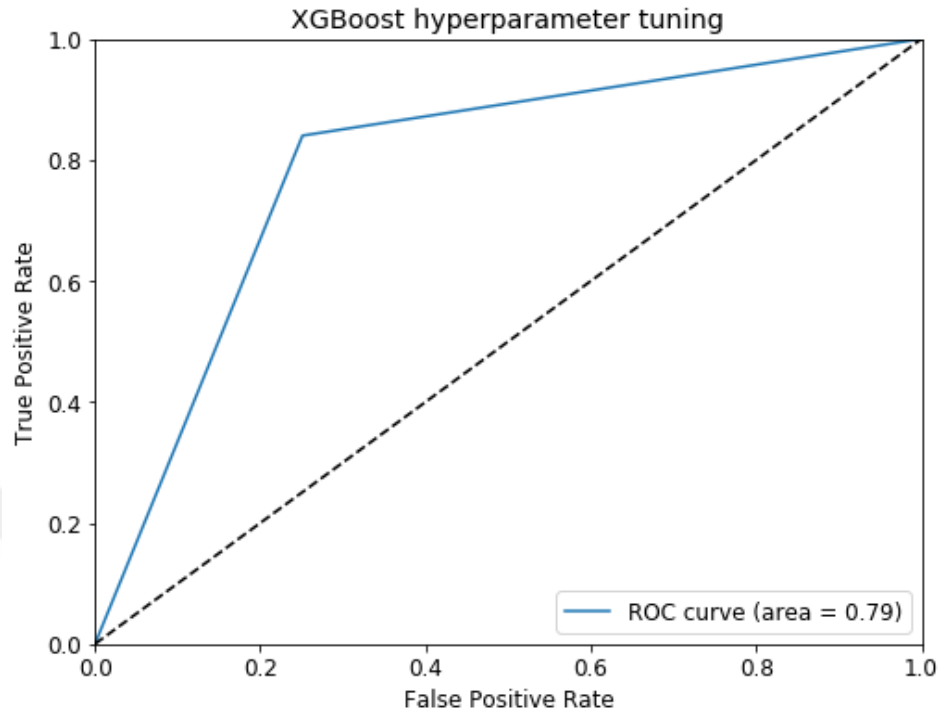

Figure 6. 9: Showed The ROC Curve of Advanced XGBoost Model Before SMOTE**.**

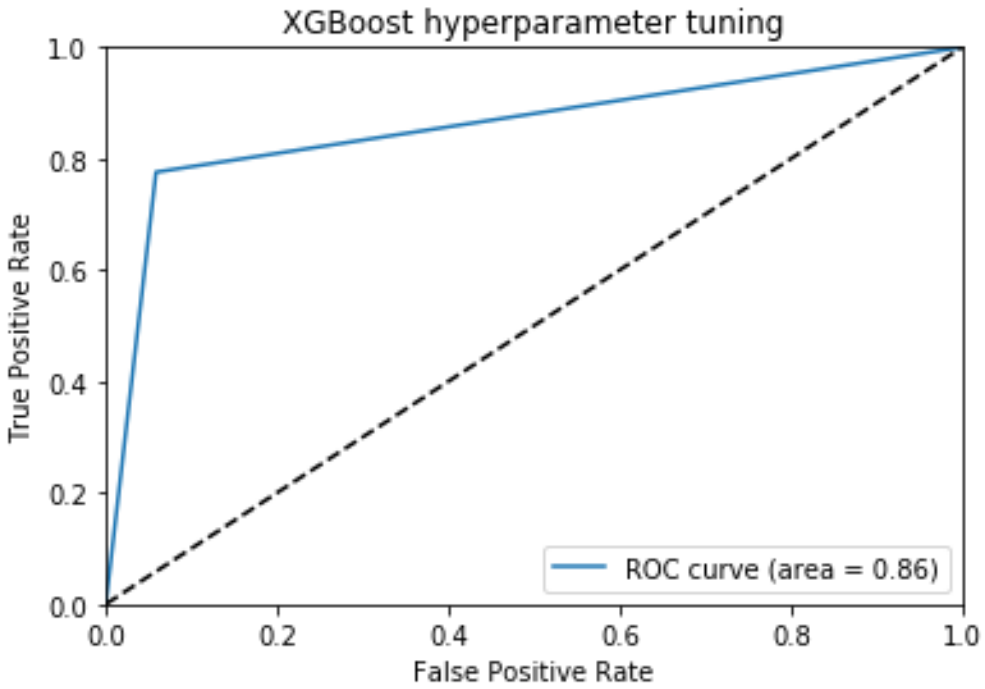

Figure 6. 10: Showed The ROC Curve of Advanced XGBoost Model After SMOTE**.**

6.1.6 Advanced Ensemble Bagging Model

We used advanced ensemble bagging model in order to attain advanced high performance model in this segment. With 49331 CVD data reports, with the better bagging model, we got 82 percent greater precision. In table 6.6, we presented the findings of the advanced ensemble bagging method with model form and accuracy.

Table 6. 8: The results of advanced ensemble bagging model with resampled CVD data.

Model Type	**Model Accuracy**
Normal Random Forest Classifier	0.8205 (+/- 0.0077)
Bagging Random Forest Classifier	0.8211 (+/- 0.0070)
Normal Extra Trees Classifier	0.8147 (+/- 0.0031)
Bagging Extra Trees Classifier	0.8204 (+/- 0.0059)
Normal K-Neighbors Classifier	0.8190 (+/- 0.0055)
Bagging K-Neighbors Classifier	0.8187 (+/- 0.0048)
Normal SVC Classifier	0.7958 (+/- 0.0090)
Bagging SVC Classifier	0.7963 (+/- 0.0093)

6.1.7 Advanced AdaBoost Model

We implemented AdaBoost hyper-parameter tuning in order to achieve a high performance model in this segment. Via a uniformly spaced grid called Grid Search (GS) including the most powerful search space exploration technique. We received 80 percent improved precision with 49331 of CVD data records with the strongest parameter set on creation including ('learning_rate': 0.1, 'n_estimators': 200) with 49331 of CVD data records. The results of the classification report of the configured advanced AdaBoost model are summarized in Table 6.7.

Table 6. 9: The results of advanced AdaBoost model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.64	0.70	0.66	4254
Cardio 1	0.87	0.84	0.86	10546
Avg.	0.80	0.80	0.80	14800

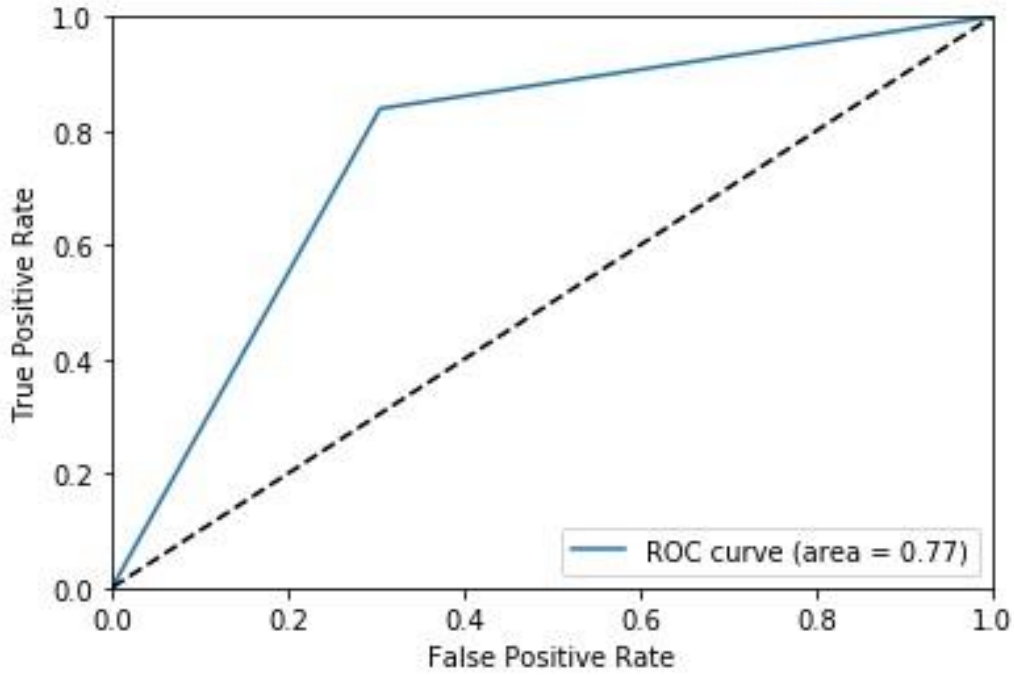

Figure 6. 11: Showed The ROC Curve of Advanced AdaBoost Model.

6.1.8 Advanced Stacking Model

In order to accomplish the advanced stacking model in this section, we introduced the advanced stacking ensemble model. With the best structured model and CVD data files, we increased accuracy by 92 percent. Table 5 shows the results of the advanced ensemble stacking approach in terms of model structure and accuracy.

Table 6. 10: The results of advanced stacking model.

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.90	0.94	0.92	10602
Cardio 1	0.94	0.89	0.91	10386
Avg.	0.92	0.92	0.92	20988

Table 6. 11: The learner structure of advanced stacking model.

Optimized Algorithm	K-Fold 0	K-Fold 1	K-Fold 2	K-Fold 3	K-Fold 4	FULL%
KNN	88.3	88	87.8	88.5	88	88
RANDOM FOREST	91.2	90.9	90.9	91	91.3	91
XGBOOST	84.7	84.3	84.8	85	84.5	84.7

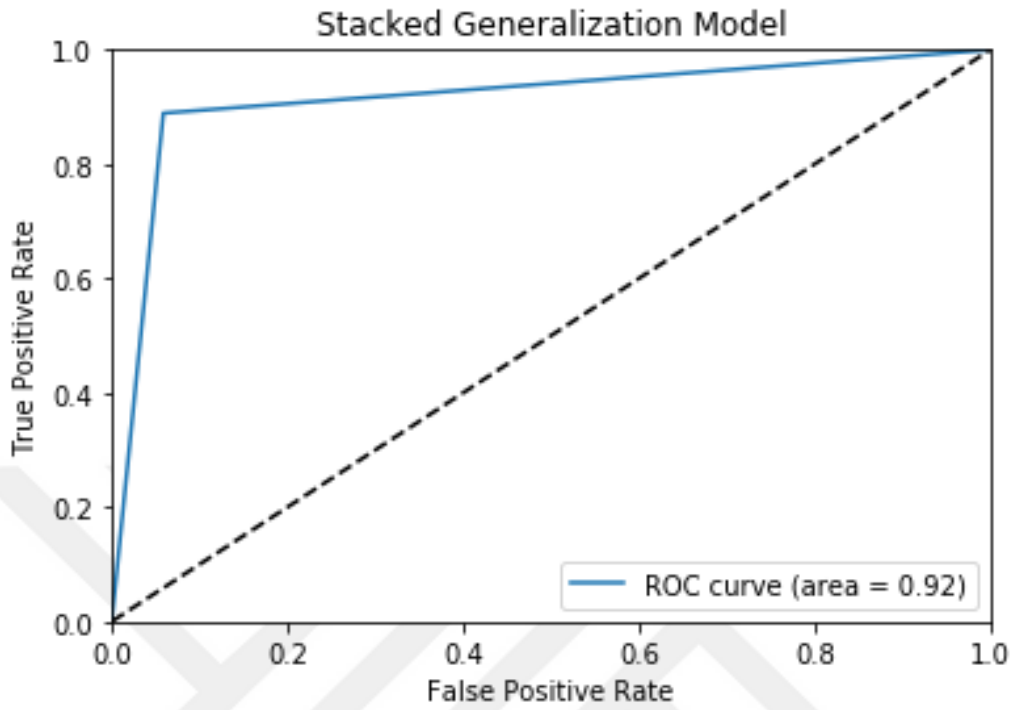

Figure 6. 12: Showed The ROC Curve of Advanced AdaBoost Model.

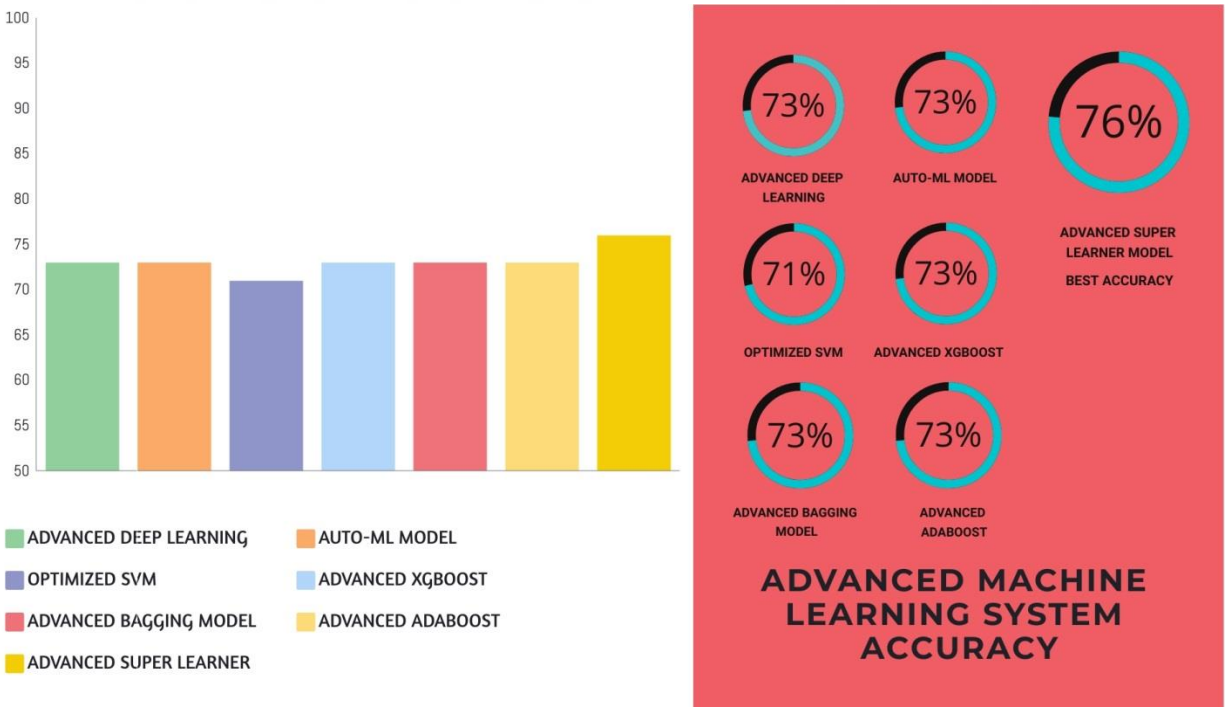

Figure 6. 13: Showed The Accuracy of Advanced Models Before Resampling Method.

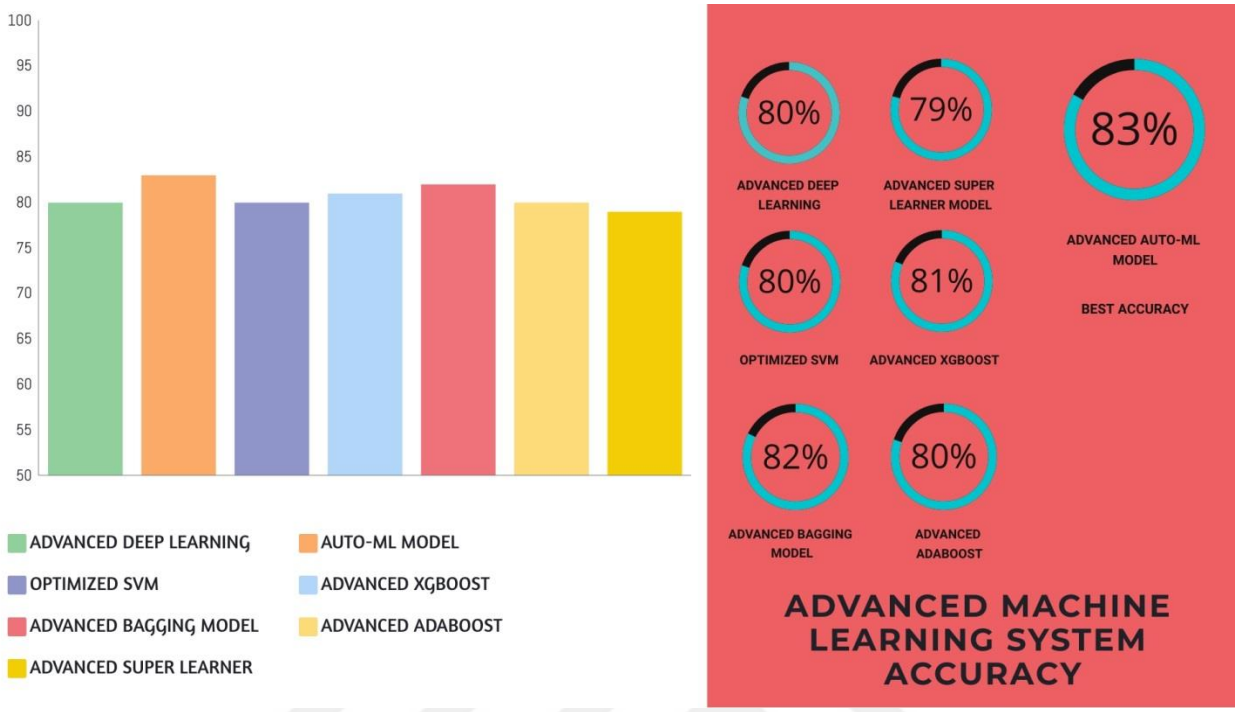

Figure 6. 14: Showed The Accuracy of Advanced Models After Resampling Method.

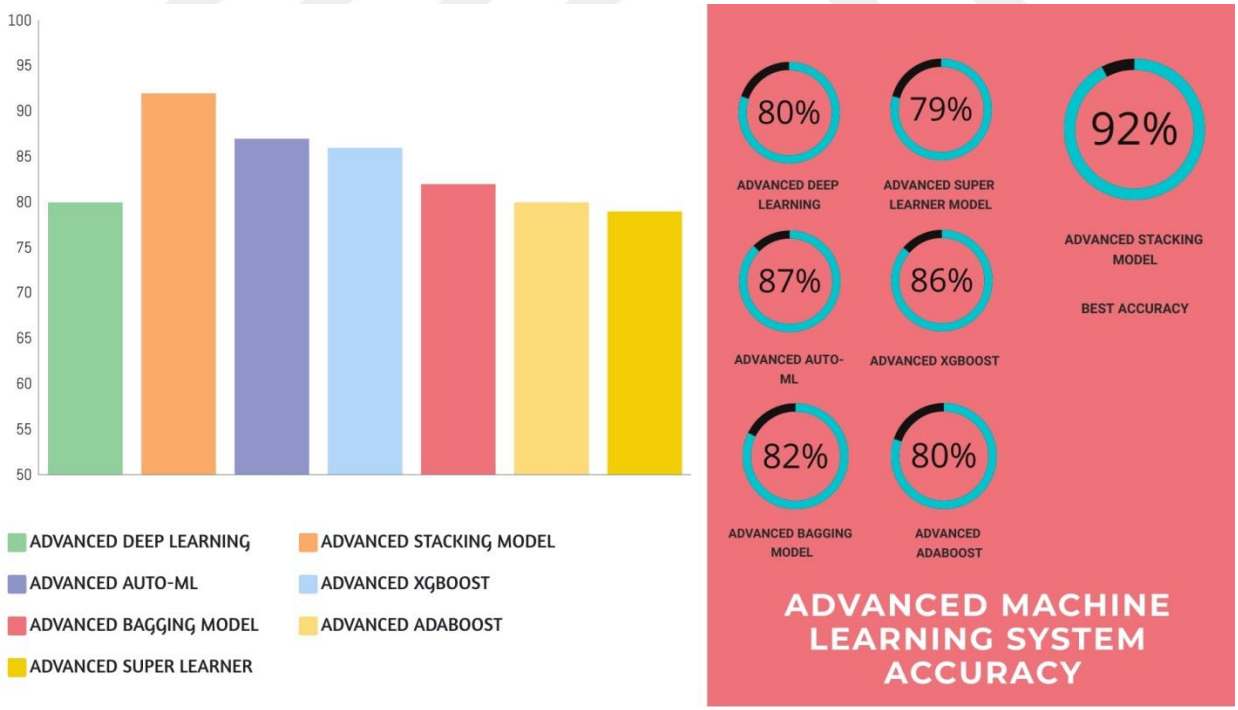

Figure 6. 15: Showed The Accuracy of Advanced Models After SMOTE Method.

6.2 THESIS SUMMARY

In 2015, at 41.5 per cent or 102.7 million, the U.S. populace had at least one CVD disease. In 2035, the number of Americans with CVD is expected to rise to 131.2 million-45 percent of the total U.S. population. In the last five years, computer learning and AI have made tremendous progress. AI techniques have gained considerable interest from clinical scientists. A survey conducted around 77 health care organizations interviewed showed that almost half of health care organizations are utilizing or intending to utilize AI in imaging. Unfortunately, it typically takes a considerable amount of time and experience of how the fundamental algorithms operate to create ML algorithms. The tuning and preparation of DNN techniques, for instance, require weeks to months. A team of human specialists and deep learning experts have developed a variety of state-of-the-art deep networks. In the biomedical and therapeutic realms, the advent of Auto-ML is theoretically revolutionary. Auto-ML may allow it by removing the high technological barrier in their practice and study; doctors will use AI methods more widely. Auto-ML can be seen as the end-to-end method on an arbitrarily defined dataset to look for the right AI model configuration. Each configuration is the product of many decisions being made on which algorithm, a form of optimization, or hyper-parameter to use. It is computationally costly to identify the right model because of the large number of configurations in the search room. Auto-ML approaches have been able to scale up significantly as computer processing technology advances due to new devices such as (GPU) and (TPU) units and increasingly powerful search algorithms. Recent research has shown that classifiers produced by computational processes have matched or even exceeded those developed by human experts. The goal is to automate developing a high-performance and sophisticated model that provides competitive outcomes on any provided dataset.

This study has dealt with Auto-ML high performance and advanced machine learning framework including advanced XGBoost, Advanced Bagging ensemble models to get a high-performance framework for applying CVD 70000 data records for the diagnosis of CVD disease. The study encourages the usage of advanced high-performance machine learning models for clinical application by breaching the assumption that ML is only open to qualified experts. For the first time, before the re-sampling process, the checked innovative and optimized machine learning algorithm with the CVD dataset. Includes Auto-ML and advanced XGBoost. Auto-ML

technology advancement is likely to allow ML instruments more available and speeds up the clinical community's science discovery process. While this research focuses on datasets of cardiovascular disorders, other biomedical datasets would maintain the main results relevant to the feasibility and efficacy of Auto-ML. This study has generated re-sampled CVD data in the second section, using the high-performance neighbourhood cleaning rule (NCL) process. This study has implemented the advanced high-performance machine learning models with re-sampled CVD data that improved the results and accuracy better than the previous analysis. Finally, the advanced machine learning technology can enable doctors and physicians in their jobs to get accurate CVD diagnosis knowledge in less time and at less expense.

6.3 COMPARISON

We compared this study against previous recent research to improve our thesis performance the table below presented the comparison using accuracy score.

Table 6. 12: The comparison of this study against previous other.

Author	Objective	Methods	Accuracy
Meghana Padmanabhan et al.[34]	A case study of risk analysis of CVD disorder.	Graduate student	74%
		Auto-ML	74%
This study with CVD before re-sampling	Diagnosis of Cardivascular CVD Disease	Advanced deep learning	73%
		Super Learner	76%
This study with CVD data after re-sampling	Diagnosis of Cardivascular CVD Disease	Advanced deep learning	80%
		Super Learner	79%
		Auto-ML	87%
		Stacking Model	92%

CHAPTER 7

CONCLUSION

This study encourages the usage of advanced high-performance machine learning models for clinical application by breaching the assumption that ML is only open to qualified experts. Firstly, before the re-sampling process the checked innovative and optimized machine learning algorithms with the CVD dataset. Includes AutoML, advanced XGBoost, advanced super learner, optimized SVM, advanced bagging models, advanced adaboost, and advanced deep learning including hyper-band hyper-parameter optimization.

AutoML technology advancement is likely to allow ML instruments more available and speeds up the clinical community's science discovery process. While this research focuses on datasets of cardiovascular disorders, other biomedical datasets would maintain the main results relevant to the feasibility and efficacy of AutoML.

This study has generated re-sampled CVD data in the second section, using the high-performance neighbourhood cleaning rule (NCL) process and advanced over-sampling SMOTE method. This study has implemented the advanced high-performance machine learning models with re-sampled CVD data that improved the results and accuracy better than the previous analysis. Finally, the advanced machine learning technology can enable doctors and physicians in their jobs to get accurate CVD diagnosis knowledge in less time and at less expense.

REFERENCES

[1] S. S. Virani *et al.*, *Heart disease and stroke statistics—2020 update: A report from the American Heart Association*. 2020.

[2] S. B. Dunbar *et al.*, "Projected Costs of Informal Caregiving for Cardiovascular Disease: 2015 to 2035: A Policy Statement From the American Heart Association," *Circulation*, vol. 137, no. 19, pp. e558–e577, 2018.

[3] S. Gandhi, W. Mosleh, J. Shen, and C. M. Chow, "Automation, machine learning, and artificial intelligence in echocardiography: A brave new world," *Echocardiography*, vol. 35, no. 9, pp. 1402–1418, 2018.

[4] H. Hacilar, Y. Gormez, B.-G. Burcu, C. Gezer, Z. Aydin, and V. C. Gungor, "Classification of Cardiovascular Artery Diseases Using Artificial Neural Network," *Int. J. Biosci. Biochem. Bioinforma.*, vol. 10, no. 1, pp. 34–41, 2020.

[5] F. Mohr, M. Wever, and E. Hüllermeier, "ML-Plan: Automated machine learning via hierarchical planning," *Mach. Learn.*, vol. 107, no. 8–10, pp. 1495–1515, 2018.

[6] C. Zhang and Z. Ye, "Water pipe failure prediction using AutoML," *Facilities*, vol. 39, no. 1–2, pp. 36–49, 2020.

[7] D. Pirova, B. Zaberzhinsky, and A. Mashkov, "Detecting heart disease symptoms using machine learning methods," *CEUR Workshop Proc.*, vol. 2667, pp. 260–263, 2020.

[8] S. Parida, S. Mallavarapu, S. Abanteriba, M. Franz, and W. Gruener, *Seating Postures for Autonomous Driving Secondary Activities*, vol. 145. Springer Singapore, 2019.

[9] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, vol. 6, no. 1, 2019.

[10] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine learning SNP based prediction for precision medicine," *Front. Genet.*, vol. 10, no. MAR, 2019.

[11] E. Walsh, "Pathological prediction: a top-down cause of organic disease," *Synthese*, 2021.

[12] "Coronary artery disease (CAD) is the most common type of heart disease and the leading

cause of death in the United States in both men and women." [Online]. Available: https://www.cormedicalgroup.com/conditions/coronary-artery-disease/. [Accessed: 12-Feb-2021].

[13] R. Alizadehsani *et al.*, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Comput. Biol. Med.*, vol. 111, no. July, p. 103346, 2019.

[14] M. P. Sendak *et al.*, "A Path for Translation of Machine Learning Products into Healthcare Delivery," *EMJ Innov.*, no. January, 2020.

[15] J. Goecks, V. Jalili, L. M. Heiser, and J. W. Gray, "How Machine Learning Will Transform Biomedicine," *Cell*, vol. 181, no. 1, pp. 92–101, 2020.

[16] D. Ben-Israel *et al.*, "The impact of machine learning on patient care: A systematic review," *Artif. Intell. Med.*, vol. 103, p. 101785, 2020.

[17] S. Kulkarni, N. Seneviratne, M. S. Baig, and A. H. A. Khan, "Artificial Intelligence in Medicine: Where Are We Now?," *Acad. Radiol.*, vol. 27, no. 1, pp. 62–70, 2020.

[18] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behav. Ther.*, vol. 51, no. 5, pp. 675–687, 2020.

[19] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020.

[20] Y. Wang *et al.*, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Inform.*, vol. 102, p. 103364, 2020.

[21] F. Guo *et al.*, "Exploration of the mechanism of traditional Chinese medicine by AI approach using unsupervised machine learning for cellular functional similarity of compounds in heterogeneous networks, XiaoErFuPi granules as an example," *Pharmacol. Res.*, vol. 160, no. June, p. 105077, 2020.

[22] M. I. Tariq, S. Tayyaba, M. W. Ashraf, and V. E. Balas, *Deep learning techniques for optimizing medical big data*. Elsevier Inc., 2020.

[23] “Deep Learning to the Rescue - Security Info Watch.” [Online]. Available: https://www.securityinfowatch.com/deep-learning-to-the-rescue. [Accessed: 14-Feb-2021].

[24] A. S. Mursch-Edlmayr *et al.*, “Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: Translation to clinical practice,” *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 1–21, 2020.

[25] D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang, and Y. Yin, “Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction,” *IEEE Access*, vol. 8, pp. 194894–194903, 2020.

[26] C. Krittanawong, H. J. Zhang, Z. Wang, M. Aydar, and T. Kitai, “Artificial Intelligence in Precision Cardiovascular Medicine,” *J. Am. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, 2017.

[27] X. Liu *et al.*, “A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method,” *Comput. Math. Methods Med.*, vol. 2017, 2017.

[28] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease,” *Proc. - IEEE Symp. Comput. Commun.*, no. Iscc, pp. 204–207, 2017.

[29] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, “Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis,” *Phys. A Stat. Mech. its Appl.*, vol. 482, pp. 796–807, 2017.

[30] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telemat. Informatics*, vol. 36, pp. 82–93, 2019.

[31] N. Louridi, M. Amar, and B. El Ouahidi, “Identification of Cardiovascular Diseases Using Machine Learning,” *7th Mediterr. Congr. Telecommun. 2019, C. 2019*, pp. 1–6, 2019.

[32] N. I. Hasan and A. Bhattacharjee, “Deep Learning Approach to Cardiovascular Disease Classification Employing Modified ECG Signal from Empirical Mode Decomposition,”

Biomed. Signal Process. Control, vol. 52, pp. 128–140, 2019.

[33] R. Sivaranjani and N. Yuvaraj, "Artificial intelligence model for earlier prediction of cardiac functionalities using multilayer perceptron," *J. Phys.*, vol. 1362, no. 1, 2019.

[34] M. Padmanabhan, P. Yuan, G. Chada, and H. Van Nguyen, "Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction," *J. Clin. Med.*, vol. 8, no. 7, p. 1050, 2019.

[35] I. Karagoz, "Prediction of Heart Diseases Using Majority Voting Ensemble Method. Cmbebih 2019," *IFMBE Proc. - C.*, vol. 73, no. May 2019, pp. 159–163, 2019.

[36] J. H. Joloudari *et al.*, "Coronary Artery Disease Diagnosis Ranking the Significant Features Using a Random Trees Model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, pp. 1–24, 2020.

[37] Z. M. Elgamal, N. B. M. Yasin, M. Tubishat, M. Alswaitti, and S. Mirjalili, "An Improved Harris Hawks Optimization Algorithm With Simulated Annealing for Feature Selection in the Medical Field," *IEEE Access*, vol. 8, pp. 186638–186652, 2020.

[38] M. Elhoseny and K. Shankar, "Optimal bilateral filter and Convolutional Neural Network based denoising method of medical image measurements," *Meas. J. Int. Meas. Confed.*, vol. 143, pp. 125–135, 2019.

[39] J. Lee *et al.*, "2017-Lee-Deep Learning in Medical Imaging_ Gen," vol. 18, no. 4, pp. 570–584, 2017.

[40] "Ulianova, S. Cardiovascular Disease Dataset." [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. [Accessed: 09-Feb-2020].

[41] J. Gatto, R. Lanka, Y. Iwashita, and A. Stoica, "Single Sample Feature Importance: An Interpretable Algorithm for Low-Level Feature Analysis," pp. 1–13, 2019.

[42] W. La Cava, C. Bauer, J. H. Moore, and S. A. Pendergrass, "Interpretation of machine learning predictions for patient outcomes in electronic health records," Mar. 2019.

[43] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive feature selection using random forest:

Selecting low-cost subsets of informative features," *Knowledge-Based Syst.*, vol. 95, pp. 1–11, 2016.

[44] "Advancing Technology Industrialization Through Intelligent Software ... - Google Books." [Online]. Available: https://books.google.com.tr/books. [Accessed: 12-Feb-2020].

[45] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2101, pp. 63–66, 2001.

[46] M. Janicka, M. Lango, and J. Stefanowski, "Using Information on Class Interrelations to Improve Classification of Multiclass Imbalanced Data: A New Resampling Algorithm," *Int. J. Appl. Math. Comput. Sci.*, vol. 29, no. 4, pp. 769–781, 2019.

[47] J. H. Seo and Y. H. Kim, "Machine-learning approach to optimize smote ratio in class imbalance dataset for intrusion detection," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[48] H. Dong, D. He, and F. Wang, "SMOTE-XGBoost using Tree Parzen Estimator optimization for copper flotation method classification," *Powder Technol.*, vol. 375, pp. 174–181, 2020.

[49] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–18, 2017.

[50] "Practical Convolutional Neural Networks: Implement advanced deep learning ... - Mohit Sewak, Md. Rezaul Karim, Pradeep Pujari - Google Books." .

[51] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, 2017.

[52] S. P. Antonio Gulli, Amita Kapoor, "Deep Learning with TensorFlow 2 and Keras," 2019.

[53] N. K. Manaswi, *Deep Learning with Applications Using Python*. 2018.

[54] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol.

18, pp. 1–52, 2018.

[55] M. Zheng, W. Tang, and X. Zhao, “Hyperparameter optimization of neural network-driven spatial models accelerated using cyber- enabled high-performance computing,” *Int. J. Geogr. Inf. Sci.*, vol. 00, no. 00, pp. 1–32, 2018.

[56] J. Wang, S. Li, Z. An, X. Jiang, W. Qian, and S. Ji, “Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines,” *Neurocomputing*, no. xxxx, pp. 1–13, 2018.

[57] “Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs ... - Antonio Gulli, Amita Kapoor, Sujit Pal - Google Books.” [Online]. Available: https://books.google.com.tr/books. [Accessed: 12-Feb-2021].

[58] S. Wu *et al.*, “L1 Norm Batch Normalization for Efficient Training of Deep Neural Networks,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 7, pp. 2043–2051, 2019.

[59] A. P. Keil and J. K. Edwards, “You are smarter than you think: (super) machine learning in context,” *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 437–440, 2018.

[60] A. Naimi, “University of Massachusetts Amherst Stacked generalization : an introduction to super learning Stacked generalization : an introduction to super learning,” 2018.

[61] R. Pirracchio and M. Carone, “The Balance Super Learner: A robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching,” *Stat. Methods Med. Res.*, vol. 27, no. 8, pp. 2504–2518, 2018.

[62] M. A. Ladds, A. P. Thompson, J. P. Kadar, D. Slip, D. Hocking, and R. Harcourt, “Super machine learning: Improving accuracy and reducing variance of behaviour classification from accelerometry,” *Anim. Biotelemetry*, vol. 5, no. 1, pp. 1–9, 2017.

[63] L. Ali *et al.*, “An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure,” *IEEE Access*, vol. 7, no. c, pp. 54007–54014, 2019.

[64] J. R. Askim, Z. Li, M. K. LaGasse, J. M. Rankin, and K. S. Suslick, “An optoelectronic nose for identification of explosives,” *Chem. Sci.*, vol. 7, no. 1, pp. 199–206, 2016.

[65] H. A. Fayed and A. F. Atiya, “Speed up grid-search for parameter selection of support vector machines,” *Appl. Soft Comput. J.*, vol. 80, pp. 202–210, 2019.

[66] A. Rojas-Dominguez, L. C. Padierna, J. M. Carpio Valadez, H. J. Puga-Soberanes, and H. J. Fraire, “Optimal Hyper-Parameter Tuning of SVM Classifiers with Application to Medical Diagnosis,” *IEEE Access*, vol. 6, pp. 7164–7176, 2017.

[67] R. Maipradit, H. Hata, and K. Matsumoto, “Sentiment Classification using N-gram IDF and Automated Machine Learning,” pp. 10–13, 2019.

[68] F. Fabris and A. A. Freitas, “Analysing the Overfit of the Auto-sklearn Automated Machine Learning Tool,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11943 LNCS, pp. 508–520, 2019.

[69] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters,” *IEEE Access*, vol. 7, pp. 118310–118321, 2019.

[70] Y. Han *et al.*, “Coupling a Bat Algorithm with XGBoost to Estimate Reference Evapotranspiration in the Arid and Semiarid Regions of China,” *Adv. Meteorol.*, vol. 2019, 2019.

[71] R. Sikora and O. Al-Laymoun, “A modified stacking ensemble machine learning algorithm using genetic algorithms,” *Artif. Intell. Concepts, Methodol. Tools, Appl.*, vol. 1, no. 1, pp. 395–405, 2016.

[72] S. E. Roshan and S. Asadi, “Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization,” *Eng. Appl. Artif. Intell.*, vol. 87, no. May 2019, p. 103319, 2020.

[73] S. F. Shetu, M. Saifuzzaman, N. N. Moon, S. Sultana, and R. Yousuf, *Student’s performance prediction using data mining technique depending on overall academic status and environmental attributes*, vol. 1166. 2021.

[74] A. Desai and S. Chaudhary, “Distributed AdaBoost Extensions for Cost-sensitive Classification Problems,” *Int. J. Comput. Appl.*, vol. 177, no. 12, pp. 1–8, 2019.

[75] V. J. Kadam and S. M. Jadhav, “Performance analysis of hyperparameter optimization methods for ensemble learning with small and medium sized medical datasets,” *J. Discret. Math. Sci. Cryptogr.*, vol. 23, no. 1, pp. 115–123, 2020.

[76] R. Gao and Z. Liu, “An Improved AdaBoost Algorithm for Hyperparameter Optimization,” *J. Phys. Conf. Ser.*, vol. 1631, no. 1, 2020.