DETECTION OF CHRONIC DISEASES USING DEEP VERSUS MACHINE LEARNING TECHNIQUES

2023
MASTER THESIS
COMPUTER ENGINEERING

Ahmed Abbas ABD ULSADA

Thesis Advisor
Assist. Prof. Dr. Nehad T.A. RAMAHA

DETECTION OF CHRONIC DISEASES USING DEEP VERSUS MACHINE LEARNING TECHNIQUES

Ahmed Abbas ABD ULSADA

Thesis Advisor

Assist. Prof. Dr. Nehad T.A. RAMAHA

T.C.

Karabuk University

Institute of Graduate Programs

Department of Computer Engineering

Prepared as

Master Thesis

KARABUK

April 2023

I certify that in my opinion the thesis submitted by Ahmed Abbas ABD ULSADA titled "DETECTION OF CHRONIC DISEASES USING DEEP VERSUS MACHINE LEARNING TECHNIQUES" is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Nehad T.A. RAMAHA .....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. April 27, 2023

Examining Committee Members (Institutions)	Signature
Chairman : Assist. Prof. Dr. İsa AVCI (KBU)	.....
Member : Assist. Prof. Dr. Murat KOCA (YYU)	.....
Member : Assist. Prof. Dr. Nehad T.A. RAMAHA (KBU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Müslüm KUZU .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Ahmed abbas ABD ULSADA

ABSTRACT

M. Sc. Thesis

DETECTION OF CHRONIC DISEASES USING DEEP VERSUS MACHINE LEARNING TECHNIQUES

Ahmed Abbas ABD ULSADA

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Nehad T.A. RAMAHA
April 2023, 65 pages

Chronic diseases are among the most prevalent diseases around the world. Moreover, chronic diseases require ongoing treatment and management, which makes it account for the bulk of healthcare spending worldwide. Hospitalization, long-term impairment, decreased quality of life, and even death are all possible outcomes of chronic diseases. Cancer, diabetes, high blood pressure, heart disease, lung disease, and kidney ailment are just some diseases on that list. Chronic illnesses are, in point of fact, the leading cause of death and disability on a global scale. This paper presents deep learning (DL) and machine learning (ML) based models for diagnosing chronic diseases. The system consists of multiple phases, including data pre-processing and disease detection. The former relies on a deep Convolution Neural Network (CNN). At the same time, the latter is supported by five machine learning algorithms: Stochastic Gradient Descent (SGD), Naive Bayes (NB), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT). The suggested model can accurately classify disorders of the

heart, the blood vessels, and the kidneys using data from three different sources (the Pima Indians Diabetes Dataset, the Cardiovascular Disease Dataset, and the UCI Heart Disease Data). As demonstrated by the experiments, without employing data augmentation, the accuracy results for datasets 1, 2, and 3 were 94%, 79.3%, and 88.01%, respectively, while the precision results were 99.77%, 99.78%, and 99.74%. Therefore, the proposed ChronicCNN model produced the best results. For datasets 1, 2, and 3, the accuracy values were (99.81%, 99.92%, and 99.77%) while the precision results were (99.08%, 99.94%, and 99.87%). In comparison, the accuracy was 94% using the SGD and LR algorithms in the second model.

Key Words : Supervised machine learning; decision tree; convolution neural network; chronic diseases; deep learning; accuracy.

Science Code : 92432

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENİMİ TEKNİKLERİNE KARŞI DERİN ÖĞRENME KULLANARAK KRONİK HASTALIKLARIN TESPİTİ

Ahmed Abbas ABD ULSADA

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Nehad T.A. RAMAHA
April 2023, 65 sayfa

Dünya çapında en yaygın hastalık kronik hastalıklarıdır. Ayrıca, kronik hastalıklar, günümüzde sağlık harcamalarının çoğunu oluşturan sürekli tedavi ve gözetim gerektiren bir hastalıktır. Hastane yatışları, uzun süreli rahatsızlanmalar, yaşam kalitesinin düşmesi ve hatta ölümler kronik hastalıkların olası sonuçlarıdır. Kanser, diyabet, yüksek tansiyon, kalp hastalığı, akciğer hastalığı ve böbrek hastalıkları bu bozuklukların yaygın örneklerindendir. Kronik koşullar küresel ölçekte ölüme ve sakatlığa yol açmaktadır. Bu çalışma müzmin hastalıkların doğru bir şekilde teşhis edilmesi için derin öğrenme (DÖ) ve makine öğrenme (MÖ) tabanlı modeller sunmaktadır. Önerilen sistem, veri ön işleme ve hastalık tespiti dahil olmak üzere birçok aşamadan oluşmaktadır. Bunlardan ilki, derin Evrişimli Sinir Ağına (ESA) dayanırken, ikincisi, hastalık tespiti için Stokastik Gradyan İniş (SGİ), K-En Yakın Komşu (KEYK), Naive Bayes (NB), Lojistik Regresyon (LR) ve Karar Ağacı (KA) olmak üzere beş makine öğrenim algoritması tarafından desteklenmektedir. Önerilen

model, Pima Kızılderelileri Diyabet Veri Kümesi, Kardiyovasküler Hastalık Veri Kümesi ve UCI Kalp Hastalığı Veri Kümesi olmak üzere bu üç farklı kaynaktan alınan verileri kullanarak kalp, kan damarı ve böbrek bozukluklarını sınıflandırmada önemli tanısal doğruluğa sahiptir. Deneysel sonuçlara dayanarak, önerilen kronikCNN modeli, en yüksek doğruluk oranıyla diğer modellerden daha üstün bir performans sergiledi. Veri artırma olmadan, 1, 2 ve 3 veri setleri sırasıyla %94, %79.3 ve %88.01 doğruluk oranı sağlandı ve %99.77, %99.78 ve %99.74 kesinlik sonuçları elde edildi. Ancak, önerilen kronikCNN modeli uygulandığında, doğruluk oranı önemli ölçüde yükselerek sırasıyla %99.81, %99.92 ve %99.77 olarak gerçekleşirken hassasiyet değerleri de %99.08, %99.94 ve %99.87 olmuştur. Karşılaştırmalı olarak, SGD ve LR algoritmalarını kullanan ikinci model sadece %94 doğruluk oranı vermiştir.

Anahtar Kelimeler: Denetimli makine öğrenimi; karar ağacı, Evrişimsel sinir ağları; kronik hastalıklar; derin öğrenme.

Bilim Kodu **:**92432

ACKNOWLEDGMENT

I am deeply grateful to my advisor Assist. Prof. Dr. Nehad T.A. RAMAHA, for his unwavering support and guidance throughout my master's program. His expertise and patience have been invaluable to me and have played a crucial role in the success of this thesis.

I am grateful to the University of Karabuk for providing me with the opportunity to conduct my research and for all of the resources and support they provided.

I am deeply thankful to my friends and family for their love and support during this process. Without their encouragement and motivation, I would not have been able to complete this journey.

Finally, I would like to extend my sincere gratitude to all of the participants in my study. Their willingness to share their experiences and insights has been invaluable to my research and has helped to make this thesis a success. Thank you for your time and contribution.

I am grateful to everyone who has supported me throughout this process. Without your help and guidance, this thesis would not have been possible.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABBREVIATIONS

1D	: 1 Dimensional
2D	: 2 Dimensional
AI	: Artificial Intelligent
CNN	: Convolution Neural Network
DL	: Deep Learning
LDA	: Linear Discriminant Analysis
ML	: Machine Learning
WHO	: World Health Organization
SGD	: Stochastic Gradient Descent
NB	: Naïve Bayes
KNN	: K-Nearest Neighbor
LR	: Logistic Regression
DT	: Decision Tree

PART 1

INTRODUCTION

1.1. OVERVIEW

According to the World Health Organization [1], the four most frequent forms of chronic illness are those that affect the digestive system, the respiratory system (such as COPD and asthma), the cardiovascular system (such as heart attacks and strokes), and the skin. Around sixty percent of all deaths that occur each year are attributable to chronic diseases, making them the leading cause of death on a global scale. As mentioned by World Health Organization [2], the leading causes of death worldwide in 2008 were cardiovascular disease, diabetes, chronic lung diseases, and cancer, accounting for 36 million deaths. However, in the next two decades, the prevalence of chronic diseases is expected to rise dramatically, with a disproportionate impact on low-income countries, populations, and communities [3].

Types of chronic diseases include heart, diabetes, and kidney diseases. One to two percent of the population is impacted by the heart failure epidemic spreading around the globe. There is a wide range of variety in both the factors that cause heart failure and its symptoms. Injuries to the heart, such as myocardial infarction, elevated preload, or afterload, produce alterations in cellular, structural, and neurohumoral pathways, impacting phenotypic characteristics. To select the treatment that will be most beneficial for each patient, it is necessary to understand the underlying pathophysiology of heart failure. In addition, removing cardiovascular risk factors is necessary to decrease the possibility of developing heart failure [4].

Diabetes mellitus is an ongoing metabolic condition that has a convoluted explanation for its pathophysiology. Problems with insulin production, its function in the body, or both can lead to hyperglycemia, which is another name for high blood sugar. The

metabolism of carbohydrates, proteins, and fats can all be adversely affected by hyperglycemia in various ways. Chronic hyperglycemia is one of the major contributors to the development of diabetic complications, both microvascular and microvascular. These consequences are the primary source of morbidity and mortality in diabetes-related illnesses. The presence of hyperglycemia is also an essential part of the diagnostic process for diabetes [5].

Living with chronic renal disease can lead to several effects, one of which is changes in the structure and function of the kidneys. Chronic kidney disease is defined as a deterioration in kidney function, an estimated glomerular filtration rate (eGFR) of less than 60 mL/min per 173 m2, or evidence of kidney damage that has been present for at least three months, such as albuminuria, hematuria, or laboratory or imaging abnormalities. Ten percent of the world's population is affected by chronic renal sickness, responsible for half a million deaths annually and the reduction of 28 million years of life. Chronic renal disease is expected to overtake cardiovascular disease as the world's sixth-biggest cause of mortality by 2040. The root causes of chronic kidney disease are not uniform across the world. However, the precise cause of chronic kidney disease is unknown, although diabetes, glomerulonephritis, and cystic kidney abnormalities are all contributing factors. Although it is unclear which comes first, hypertension and chronic renal illness are often found together [6, 7].

1.2. MOTIVATION OF THESIS

Chronic diseases linked to non-communicable diseases are long-lasting and typically not fully curable. The four main chronic illnesses include cancer, cardiovascular diseases (CVDs), congestive heart failure (CHF), cardiac arrhythmias, pulmonary circulation problems, and diabetes mellitus (e.g., type 1 diabetes, type 2 diabetes, pre-diabetes, and gestational diabetes). The frequency of these chronic diseases is rising, and they are the leading causes of death in the majority of nations around the world. According to the World Health Organization (WHO), 71% of all deaths worldwide occurred due to chronic diseases in 2016, killing 40.5 million people.

Machine learning classifiers have become crucial tools for making decisions in the risk assessment of chronic diseases. The risk of chronic diseases has been evaluated in the medical field using several machine and deep learning classifiers. The chronic risk prediction model is based on medical datasets, but most medical datasets do not have a uniform distribution and frequently include more occurrences of one class than another. The dataset becomes unbalanced as a result. Classifiers provide high accuracy for most classes and lower prediction accuracy for a few classes when trained on an unbalanced data set. The primary motivation of this work is to find a set of classifiers with better accuracy and to identify the impact of this accuracy on the results when the data are balanced and unbalanced.

1.3. PROBLEM STATEMENT

Clinicians usually manually check patients' medical histories to identify chronic diseases and draw inferences, which may be time-consuming and error-prone. Care costs and human error risk rise as medical professionals and care teams spend more time manually identifying chronic diseases from the clinical information rather than engaging with patients, delivering individualized care, and developing effective treatment plans. A computer system for identifying and simulating chronic diseases could cover this gap. Access to an intelligent disease modeling framework with detection and tracking tools would be of the utmost value. These tools would give physicians and patients a second opinion on the disease condition, as it is difficult to interpret the disease from routine clinical summary notes, visual inspection, or by reading the records. Given the seriousness of chronic diseases and the maximum number of deaths due to them worldwide, it is necessary to think about building an intelligent system that helps doctors and specialists classify these diseases. Therefore, this thesis proposes using deep learning to build a novel Convolutional Neural Network (CNN) model. Before starting the study, we reviewed the available previous studies in this field, but they were insufficient. Therefore, we suggested a new model to obtain better results in this thesis.

1.4. AIMS AND OBJECTIVES

The main objective of this research is to determine the most effective way to classify chronic diseases. This stage includes obtaining the best technology for building an intelligent system using deep learning and machine learning to build a new model of the central neural network (CNN). We may specifically emphasize the following objectives can be described as follows:

- To compare and contrast a broad range of machine-learning strategies and algorithms for the early detection of certain persistent health conditions.
- To compare the results of deep learning and machine learning in the case of data augmentation and data without augmentation.
- To compare the best results produced using deep classification for disease diagnosis with pre-trained results for disease diagnosis from the available literature.

1.5. CONTRIBUTION OF THESIS

The main contributions of this thesis are:

- Proposing a novel Convolutional Neural Network (CNN) model for the classification of chronic diseases.
- The proposed deep learning model was used in addition to a set of machine learning algorithms, and the obtained results were compared to determine the best way to classify diseases.
- Proposing a model that trains three different data sets between a huge dataset of 70,000 samples and a data set containing 80 samples and tries to maintain high accuracy and raise the precision value. In addition, this data is unbalanced. Then, we use the augmentation operation to balance the data and test the proposed model again.
- The proposed system is high-speed as it trains each epoch in 2 seconds.
- Analyzing and processing data in the proposed system is very short.

PART 2

LITERATURE REVIEW

Interest in healthcare and its development is ever-present because of its centrality to modern society and its direct impact on people's health. Many researchers have looked at the medical applications of AI methods including machine learning (ML) and data mining [8]. The healthcare industry has recently had considerable success applying artificial intelligence (AI) approaches to the exploration of knowledge from health data. Specifically, data mining methods have been widely applied to both disease diagnosis and prognosis [9].

Because early identification of chronic disease is a significant issue, numerous deep learning and machine learning techniques that are currently available have been created in order to detect a particular chronic disease. In this part of the article, a diagnosis of diabetes, heart disease, and renal disease was analyzed by looking at the results of several studies [10].

Diabetes is a condition in which the body has problems properly metabolizing glucose. This leads to hyperglycemia, or an abnormally high blood glucose level [11]. If cells in the body are unable to respond to the insulin that is generated, or if the body fails to produce enough insulin, the result is diabetes. Since there is currently no treatment for diabetes, proper management is essential. Type 1, type 2, and gestational diabetes are the three types of diabetes, with type 2 being the most common [12].

Madan et al. [13] built a hybrid deep learning system to monitor in real-time for signs of Type 2 diabetes mellitus using data from the PIMA Indian diabetes dataset. There are four main contributions to this work. CNN-Bi-LSTM was proposed as a way to predict and improve Type 2 diabetes diagnosis.

These results demonstrate that CNN-Bi-LSTM achieves an accuracy rate of 98% and precision rate of 87% than competing deep learning algorithms.

Singla et al. [14] anticipated diabetes by utilizing the same analytical process that was used to select the most essential features. In accordance with the proposed methodology, the findings point to a significant connection existing between diabetes, BMI, and glucose level. Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), and The I Bayes (NB) Methods were utilized to make an accurate diabetes prognosis. The results yielded accuracy of 80% the highest accuracy when using the ANN algorithm.

Chang et al. [15] proposed an Internet of Medical Things (IoMT) environment e-diagnosis system that is dependent on machine learning (ML) methodologies. The system was designed primarily for identifying diabetes mellitus (type 2 diabetes). The diabetes dataset from the Pima Indians will be used to train and assess three interpretable supervised machine-learning models. These interpretable models include the J48 decision tree models, the Naive Bayes (NB) classifier, and the random forest classifier. An NB model works well with a more limited collection of data, with accuracy rate of 79.13% and 81.88% of precision, the J48 with accuracy rate of 75.22% and 70.86% of precision rate whereas a Random Forest (RF) model performs better with more features, having 89.40% accuracy rating and 75.22% of precision. Both models are used for binary classification.

Barik et al.[16] offers a pair of machine-learning approaches for predicting diabetes. The first strategy employs a combination of both classification and clustering techniques, whereas the second is based only on clustering. The random forest technique was chosen for this reason. For our hybrid approach, we decided on the XGBoost algorithm. The average score for these two algorithms was 74.10%, making them superior to the Random Forest approach when evaluating the accuracy of predictions for diabetes using two alternative machine learning methodologies.
Spoorthy and Sunitha [17] Predicting diabetes using a dataset using machine learning classification and ensemble approaches. Using practical classifiers like Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and XGBoost, we may

reach high levels of accuracy. On the PIMA dataset, RF achieves the best accuracy of 94.07%, making it the gold standard for diabetes diagnosis.

Khaleel and Al-Bakry [18] demonstrated a model for predicting whether or not a person will get diabetes. The accuracy, recall, and F1-measure of the predictions made by a handful of high-powered ML algorithms form the basis of the suggested model. Diabetic onset was predicted using the Pima Indian Diabetes (PIDD) dataset. The precision rates achieved by applying the Logistic Regression (LR), Nave Bayes (NB), and K-nearest Neighbor (KNN) algorithms were 94%, 79%, and 69%, respectively. The results show that LR is superior to other algorithms in its capability to predict diabetes.

Zhou et al. [19] hypothesized that a model is constructed primarily by accessing the network's hidden layers, and dropout regularization is employed to prevent overfitting. They employed the deep neural network's binary cross-entropy loss function and a few other fine-tuned parameters to make accurate predictions. The experimental outcomes validate the effectiveness and viability of the proposed DLPD model for diabetes prediction. Maximum accuracy during training is 94.02% for the general diabetes dataset and 99.4% for the Pima Indian diabetes dataset.

Challa and Chinnaiyan[20] created two machine-learning models that can accurately predict blood pressure, body mass index, and Glucose Levels (GL). The increased accuracy and transparency of the results made available by these data sets have significant implications for medical care. The proposed method achieved accuracy rate of 78.25% meets the requirements of healthcare environments for sound decision-making.

Choudhury and Gupta [21] used a variety of algorithms to divide people into high-risk and low-risk groups to categorize them. The SVM statistical machine learning technique was used to create classifiers for DTs, RF, and NB. The KNN classification approach was used to cluster newly collected data. According to the findings of this research, the NB method produced the highest precision with 85.27% precision rate,

while the LR approach was found to be more accurate and efficient with 77.61% accuracy rate.

Aminah and Saputro[22] built a method for iris-based diabetes prediction using machine learning. The system is equipped with a set of tools for taking and editing photographs. The iris was photographed with a camera mounted on an iridoscope. Feature extraction processes typically use the Gray Level Co-Occurrence Matrix (GLCM) technique to learn about a picture's texture. The KNN (k Nearest Neighbor) approach is used to divide cases of diabetes into those that do not involve the disease. A k-fold cross-validation method and a confusion matrix are then utilized to verify the accuracy of the classification findings. Experiments involved both diabetics and non-diabetics. The results show a precision of 0.889, a sensitivity of 0.796, a false-positive rate of 11.0, and a false-negative rate of 20.0.

Yahyaoui et al. [23] compared deep learning and machine learning diabetes prediction techniques. All test iterations showed that RF was superior at classifying diabetes, with a total accuracy of 83.67% for diabetic prediction. Prediction accuracy between SVM and DL on our dataset was 65.38 and 76.81 respectively. Swapna et al. [24] employed ECG signal-derived HRV information, and diabetes might be accurately diagnosed. This study uses deep learning to make a diabetes diagnosis using HRV data. Thus far, the CNN-LSTM network with 5-fold cross-validation has shown the most promise for automated diabetes detection utilizing HRV. The results yielded precision rate of 88.9% and 90.9% of accuracy.

Swapna et al. [10] extracted complex temporal dynamic features from the HRV data using Convolutional Neural Networks (CNN), long short-term memory (LSTM), and their combinations. These features are supplied into a Support Vector Machine for classification (SVM). The proposed categorization approach can help doctors predict diabetes using ECG data with a 95.7% accuracy rate. Cardiovascular disorders are a leading cause of death worldwide (CVD). Atherosclerosis is the leading cause of cardiovascular disease, and it is caused by the buildup of cholesterol and fat in the artery walls (CVD)(22). If blood flow is impeded, the entire body suffers the consequences. Atherosclerosis is a leading cause of cardiovascular disease, including

hypertension, coronary artery disease, and stroke (such as stroke). Heart failure, cardiac arrhythmias, congenital heart disease, rheumatic heart disease, and inflammatory heart disease are all examples of other cardiovascular diseases [25]. In order to address this problem, CAD software can aid doctors in making early diagnoses of illness. Different disease-related issues are predicted and investigated using CAD, and machine learning approaches are employed to do so. When it comes to analyzing large amounts of patient data, the deep learning-based technique is among the most cutting-edge technologies available. Deep learning-based algorithms improve in accuracy and effectiveness as more data is used as input.

Acharya et al. [26] proposed a CNN model that requires minimal preprocessing of the raw ECG signal and does not necessitate the use of artificial features or classification. Four different datasets were utilized for training and testing the proposed CNN model (A, B, C, and D). With an accuracy of 98.97%, a specificity of 99.01%, and a sensitivity of 98.87%, Set B had the best performance metrics among the four sets. The proposed CNN model has the potential to offer cardiologists a more objective and quick interpretation of ECG signals, making it a useful diagnostic aid.

Alqahtani et al.[27] introduced a deep learning and machine learning ensemble strategy for predicting the risk of cardiovascular disease. Their projections for cardiovascular illness are based on these six categories. A dataset of individuals with cardiovascular disease is made accessible for the training of the models. The authors use Random Forest to identify risk factors for cardiovascular disease (RF). The experiment results show that the ML ensemble model is the most effective in its ability to forecast the occurrence of a disease (88.02%) of precision and (88.70) of accuracy.

Sharifrazi et al. [28] constructed a Convolutional Neural Network-Clustering deep learning model for myocarditis detection (CNN-KCL). In this study, 98,898 images from 47 patients were analyzed for signs of myocarditis. The research indicates that the proposed approach for diagnosing myocarditis has a 97.6% precision and 97.41% accuracy rates when using 10 fold-cross validations with 4 distinct clusters.

Hussain et al. [29] presented a deep learning architecture employing a 1D convolutional neural network for distinguishing healthy and unhealthy persons, circumventing the constraints of previous systems. Patients' risk levels are evaluated using a variety of clinical criteria, which ultimately leads to earlier diagnosis. The proposed network employs many methods to guard against overfitting. Using the suggested network, we can achieve a precision of over 94.73% and accuracy reached 97.41% on the dataset. Mehmood et al. [30] proposed the Cardio Help approach, which employs the deep learning algorithm Convolutional Neural Networks to assess a patient's likelihood of developing cardiovascular disease (CNN). With a focus on temporal data modeling, the suggested method makes use of CNN for preclinical HF prediction. With an accuracy of 97% success rate on the collected dataset, experimental outcomes reveal that the proposed strategy excels above state-of-the-art methods.

X. Zhang et al. [31] Convolutional neural networks (CNNs) were trained using 259,789 ECG signals obtained from cardiac function rooms at a tertiary care hospital to identify the presence of cardiovascular disease. The CNN classification was validated using an external test dataset consisting of 18,018 ECG signals. Overall, the model's diagnostic precision was estimated to be 60.93%, with an accuracy of 95% for distinguishing atrial fibrillation from a normal cardiac rhythm. By minimizing the incidence of incorrect and missed diagnoses in primary care settings, the suggested CNN approach has the potential to increase productivity and decrease labor costs for big general hospitals. Shankar et al. [32] a convolutional neural network approach was proposed to be used on both structured and, maybe, unstructured patient data. The developed model has an accuracy of 97%. We've also proposed to include many machine learning methods in the training set, evaluating their performance, and inferring the most accurate one based on the data's overall use for illness risk prediction. Further precision can be achieved by adjusting the properties.

Singha et al. [33] proposed the use of Convolutional Neural Networks as a means of early medical diagnosis and prognosis (CNNs). 13 medical elements are fed into CNN. The convolutional neural network (CNN) is trained by a variant of the backpropagation method. Testing shows that CNN can accurately predict the presence or absence of

heart disease with a rate of precision greater than 97.70% and accuracy rate of 95%. For the most part, people with chronic kidney disease (CKD) don't notice anything is wrong until their kidney function drops to about 15–20% of normal. The medical establishment has a hard time spotting CKD early on and treating it when it first appears. Current medical professionals and academics are keenly interested in the creation of diagnostic and prognostic tools for a wide range of diseases, particularly those that are common in the human population [34]. Here we will examine the state of the art in predicting the likelihood of developing CKD using machine learning.

Mondol et al. [35] employed a total of 24 attributes in a binary CKD classification project. Traditional convolutional neural network (CNN), artificial neurologist (ANN), and long short-term memory (LSTM) models were used, as were their optimized counterparts (optimal CNN [OCNN], ANN [OANN], and LSTM [OLSTM]). Of the reference models, CNN had the highest validation accuracy 98.75%, 98.5%, and 96.25% for OCNN, OANN, and OLSTM respectively, while the precision by OCNN (96.55%), OANN (90.32%), and OLSTM (93.33%).

Al-Moman et al.[36] aimed at providing an early diagnosis of CDK by the use of machine learning methods including ANNs, SVMs, and k-Nearest Neighbors (KNN). The significance of AI is shown in the fact that the discovery of such often lethal diseases is a need. There are 400 samples and 13 variables analyzed in this study. The efficacy of the three approaches was determined by testing them on the collected data.

Ilyas et al. [37] created and analyzed two algorithms, J48 and random forest, to forecast CKD stages. J48 has the best precision and accuracy with 64% and 85.55% respectively and success rate in identifying instances. While Random forest requires 0.28 s, J48 just requires 0.03. Since J48 produces better results than Random Forest and is faster to run, it is a time- and resource-saving option. Elkholy et al[38] provide a clever approach to classifying and forecasting information. To forecast kidney-related diseases, we employ a Deep Belief Network (DBN) with a few tweaks: we use Softmax as the activation function and Categorical Cross-entropy as the loss function. The results of the evaluation show that the proposed model has a higher sensitivity of 87.5% and specificity of 98.5% than the current gold standard does for predicting

CKD. Early detection of CKD and related stages can reduce the severity of kidney damage. The results show the value of using advanced deep-learning algorithms in clinical decision-making.

Yashfi et al**.,** [39] created a strategy for risk prediction of CKD based on an examination of data from patients with CKD. We used data from 455 patients. Clinical data from Khulna City Medical College is used alongside synthetic data collected from the UCI Machine Learning Repository. Python, a general-purpose interpreted programming language, was the backbone of our system's development. The data was trained using 10-fold cross-validation (CV), in addition to Random forest and ANN. The accuracy rate reached 97.12%, whereas the precision of the Random forest approach is 97%, while that of ANN is just 94.5 % of accuracy and 94% of precision. This approach will help with the early diagnosis of chronic renal diseases.

Ghosh et al. [40]Accurate predictions were obtained by employing a variety of methods, including Support Vector Machine (SVM), AdaBoost (AB), Linear Discriminant Analysis (LDA), and Gradient Boosting (GB). The UCI machine learning repository dataset has been used to test these algorithms. The best-predicted precision, around 99% is achieved by SVM, AdaBoost, and LDA classifiers. While the best accuracy was achieved with a GB of 99.80%. After that, many metrics for gauging performance were introduced, illuminating pertinent results. Ultimately, these guidelines enable the selection of the most efficient and optimal algorithms for the given task.

Manonmani & Balakrishnan[41] applied a model to the CKD dataset using an ensemble feature selection technique. To determine which aspects of CKD are most important, researchers employ a filtering method known as density-based feature selection (DFS). In order to determine which features are most relevant for accurate CKD prediction, the DFS method's output is fed into a wrapper-based optimization process called Improved Teacher Learner Based Optimization (ITLBO). To evaluate the usefulness of the ensemble feature selection approach, this research analyzes the results obtained by using SVM, Gradient Boosting, and CNN classification strategies. The best accuracy achieved by CNN was 97.75%.

Vinothini et al, [42] used CVD data with an uneven number of positive and negative cases to train a CKD prediction model. The investigation consists of three stages: Models are initially chosen using performance criteria that take into consideration the non-uniform distribution of classes without resorting to resampling. After increasing the size of the minority class's training data set in the second stage with the Synthetic Minority Oversampling Technique (SMOTE), we then utilized random under-sampling of the majority class's training data set in the third stage to ensure statistical parity between the two groups. The data suggest that the MLP (Multi-Layer Perceptron)-SMOTE model is superior to other methods for predicting CKD. Higher F-score, recall, precision, G-mean, balanced accuracy, and RUC-AUC values are seen with this model. The MLP gives the best accuracy and precision of 93% and 64% respectively.

Kriplani et al.,[43] proposed a method for determining whether or not a person has a chronic renal disease with a success precision of 100% and accuracy of 97%. By using cross-validation to avoid overfitting, our model beats the state-of-the-art methods currently available. Early detection of chronic renal disease is crucial for receiving the life-saving benefits of automated treatment and reducing the rate of kidney damage.
Tekale et al.[44] researched numerous machine-learning approaches. Accuracy predictions for several machine learning models, such as Decision Trees (DT) and Support Vector Machines (SVM), were analyzed, and 14 unique patient characteristics related to CKD were included. The results show the precision rate of 93.08% with SVM and the Decision Tree (DT) algorithms of 85.02%. Table (2.1). Summarize relevant research that used machine learning or deep learning methods.

Table 2.1. Related literature review.

Studies using DL and ML for diabetes disease							
No.	**Year**	**Methods**	**Dataset**	**Precision**	**Accuracy**	**Recall**	**F-measure**
Madan et al. [13]	2022	CNN-Bi-LSTM	PIMA Indians diabetes	87%	98%	82%	85%
Singla et al. [14]	2022	ANN SVM DT NB	PIMA Indians diabetes	/	80%, 64.28% 74.67% 76.62%	/	/
Chang et al[15]	2022	NB RF J48 DT	PIMA Indians diabetes	81.88% 89.40% 70.86%	79.13% 75.22% 75.22%	88.08% 81.33% 89.92%	84.71% 85.17% 78.81%
Barik et al.[16]	2021	RF, and XG boost	PIMA Indian Diabetes Dataset	/	71.9%. 74.1%	/	/
Spoorthy and Sunitha [17]	2021	DT SVM RF XGBoost	PIMA Indians diabetes	/	83.54% 89.34% 92.34% 90.34%	/	/
Khaleel and Al-Bakry [18]	2021	LR NB KNN	PIMA Indians diabetes	94% 79% 69%	/	70% 68% 68%	79% 72% 68%
Zhou et al. [19]	2020	Deep Learning for Predicting Diabetes (DLPD)	PIMA Indians diabetes	/	99.4%	/	/
Challa and Chinnaiyan[20]	2020	DT SVM KNN RF	PIMA Indian Diabetes Dataset	/	78.25% 77.73% 77.47% 77.90%	65.31% 43.75% 64.10% 71.3%	/
Choudhury and Gupta[21]	2019	SVM KNN DTs NB LR	PIMA Indian Diabetes Dataset	77.50% 81.45% 80.69% 85.27% 79.79%	75.68% 75.1% 67.57% 76.64% 77.61%	89.60% 81.21% 67.63% 78.61% 89.02%	83.11% 81.33% 73.59% 81.80% 84.15%
Aminah and Saputro [22]	2019	KNN	PIMA Indian Diabetes Dataset	/	DM Prediction 85.6% Accuracy	/	/
Yahyaoui et al. [23]	2019	RF SVM DL	PIMA Indians diabetes	/	83.67% 65.38% 76.81%	/	/
Swapna et al. [24]	2018	CNN-LSTM	used 71 datasets of diabetic people and 71 datasets of normal people	88.9%	90.9%	100%	91.7%
Swapna et al. [25]	2018	CNN	PIMA Indians diabetes	/	95.7%	/	/

Studies using DL and ML for heart disease							
Acharya et al. [26]	2019	CNN	The Fantasia, BIDMC ECG, and MIT-BIH Normal Sinus databases		98.97%		
Alqahtani et al.[27]	2022	RF, KNN, DT, XGB, and two deep learning models, DNN and KDNN	Kaggle repository' s cardiovascular disease dataset	RF 88.02%	RF 88.70%	RF 88.02%	RF 88.01%
Sharifrazi et al [28]	2022	Convolutional Neural Network-Clustering (CNN-KCL)	total number of 98,898 images to diagnose myocarditis disease	97.6%	97.41%	95.7%	96.5%
Hussain et al [29]	2021	1D Convolutional Neural Network (CNN)	Cleveland database	94.73%	96%	100%	97.29%
Mehmood et al., [30]	2021	CNN	heart-disease UCI standard repository	/	97%	/	/
Zhang et al., [31]	2020	CNN	random sample of 18,018 ECGs	60.93%	95%	99.95%	/
Shankar, et al. [32]	2020	CNN	Cleveland database	/	97%	/	/
Singha et al.[33]	2018	CNNs	Cleveland database	97.70%	95%	97.62%	97.14%
Studies using DL and ML for kidney disease							
Mondol et al. [35]	2022	Optimizer Convolutional Neural Network (OCNN) Optimizer Long Short-Term Memory (OLSTM) Optimizer Artificial Neural Network (OANN)	CKD Prediction Dataset	96.55% 90.32% 93.33%	98.75% 98.5% 96.25%	98% 94% 96%	99% 97% 98%

Al-Moman et al.[36]	2022	ANN, SVM, KNN	CKD Prediction Dataset	/	ANN classifier got the highest accuracy of 99.2%.	/	/
Ilyas et al. [37]	2021	J48 Random Forest (RF)	CKD Prediction Dataset.	64% 56.1%	85.5% 78.25%	70% 36.7%	66% 56.1%
Elkholy et al[38]	2021	Modified Deep Belief Network (DBN)	CKD Prediction Dataset.	/	98.5%	/	87%
Yashfi et al., [39]	2020	RF, and ANN	UCI Machine Learning Repository	97% 95%	97.12% 94.5%	97% 94%	97% 95%
Ghosh et al. [40]	2020	SVM AdaBoost LDA Gradient Boosting	UCI Machine Learning Repository	99% 99% 99% 98%	99.56% 97.91% 97.91% 99.80%	99% 98% 98% 99%	99% 98% 98% 99%
Manonmani & Balakrishnan[41]	2020	SVM, Gradient Boosting, and CNN	CKD Prediction Dataset.	/	93% for SVM, 97% for Gradient Boosting, and 97.75% for CNN	/	/
Vinothini et al, [42]	2020	MLP KNN SVM LR	CVD dataset	64% 31% 53% 43%	93% 81% 89% 87%	56% 56% 62% 56%	60% 40% 57% 49%
Kriplani et al.,[43]	2019	CNNs	UCI Machine Learning Repository	100%	97%	95.2%	97.6%
Tekale et al.[44]	2018	DT, and SVM	CKD Prediction Dataset	85.02% 93.08%	/ /	94.66% 98.66%	89.58% 95.79%

Table 2.2. Limitation of related studies.

Ref. no.	Limitations
[13]	Using only one dataset the static PIMA Indian dataset (PIDD)
[14]	Doesn't use the process of data feature extraction
[15]	Suffering from the preprocessing stage, the accuracy may be enhanced by using suitable pre-processing techniques for data management and analysis.
[16]	It's better to use other kinds of medical data to be adapted in such a framework creating a cost-effective and time-saving option for both diabetic patients and doctors.
[17]	Using only one dataset the static PIMA Indian dataset (PIDD)
[18]	Using only one dataset the static PIMA Indian dataset (PIDD)
[19]	The amount of data that the model can handle is high. In the hyperparameter tuning method, training too many parameters can easily result in overfitting.
[20]	Using only one dataset
[21]	Using only one dataset the static PIMA Indian dataset (PIDD)
[22]	Too many pre-processors and the use of images negatively affects the time and speed of the system.
[23]	A higher number of diabetic patients were misclassified as non-diabetic patients.
[24]	Large-sized input datasets must feed into the proposed architecture compared to the dataset size used in this work.
[27]	For a better and more accurate assessment, additional datasets may be employed.
[28]	Using only one dataset
[29]	More and more parameters can be included in the system which can help in classifying heart disease more accurately.
[30]	Using only one dataset
[31]	The sensitivity and specificity of the individual classifications must be improved by adjusting the different parameters.
[32]	Less data used.
[33]	Using only one dataset
[35]	Using only one dataset
[36]	Doesn't use the process of data feature extraction
[37]	Using only one dataset
[38]	Doesn't use the process of data feature extraction
[39]	Using only one dataset
[40]	Using only one dataset
[41]	Using only one dataset
[42]	Fewer data were used.
[43]	Using only one dataset
[44]	Using only one dataset

PART 3

THEORETICAL BACKGROUND

According to the papers presented in the preceding chapter, the findings of these studies are crucial to comprehending the Chronic Diseases Classification. Additionally, they provide data for Chronic Disease Prediction. Gaining this benefit requires extensive study of Classification algorithms. This section will display these computations and knowledge and provide classification algorithms.

3.1. MACHINE LEARNING TECHNIQUES

The field of machine learning (ML) is crucial to AI since it focuses on developing self-learning algorithms for computers. To make a conclusion or uncover a pattern, ML methods can learn from data on their own without being explicitly programmed to do so. This is achieved by exposing it to numerous training sets, each of which contributes to a greater understanding of the system's underlying concept and architecture. Simply put, the algorithms are self-taught [45]. ML is defined as a program's capacity to make reliable forecasts in light of existing data. Recent years have witnessed major advancements in ML thanks to the exponential growth in computer storage space and processing capacity. One of the many benefits of ML is that it can analyze massive data sets to find patterns and correlations. The simple processing of data based on images aids specialists in making tough choices. Plus, it allows for the instantaneous processing of massive volumes of data, something the human brain just cannot do [46] Medical care is only one of several industries that employ ML methods. Given the high stakes and high costs of clinical data analysis, ML methods have been adapted for use in the healthcare industry. When development time and cost are crucial, when the topic looks too complex to study in its totality, and so on, ML can be superior to more traditional approaches [47].

Figure (3.1) shows that there are many subclasses of ML, but the big three are supervised learning, unsupervised learning, and reinforcement learning.

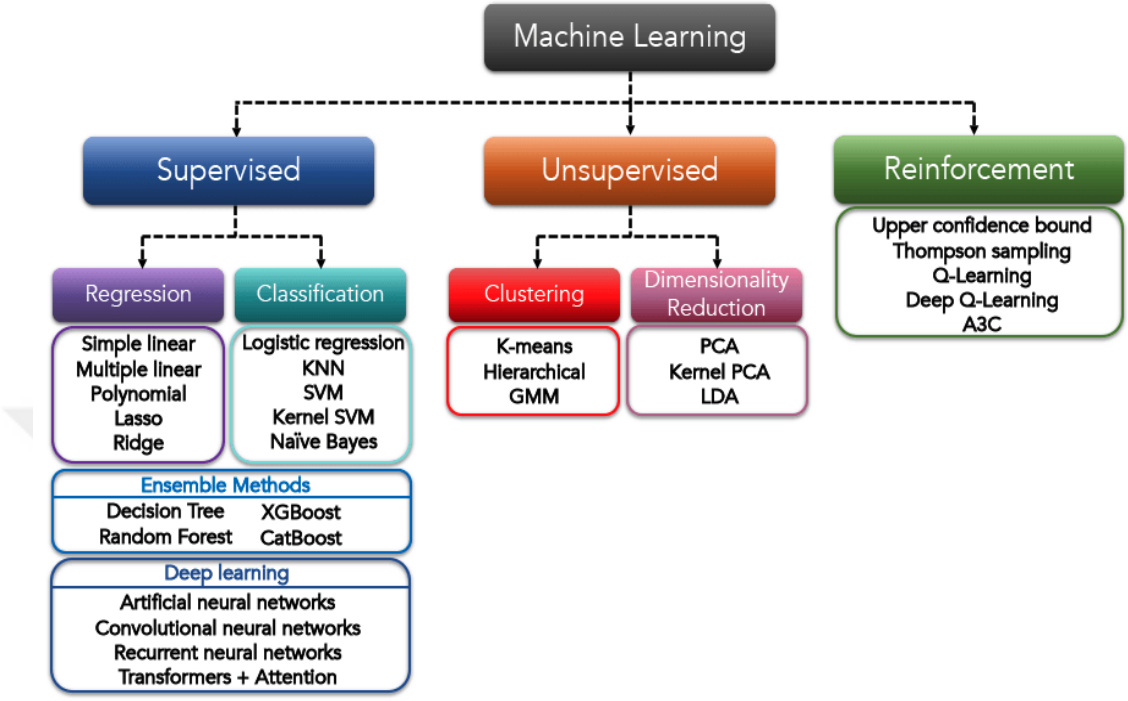

Figure 3.1. The three main categories of machine learning methods.

Supervised learning is the foundation of machine learning, where algorithms are taught to make predictions by being given inputs and outputs (features and targets, respectively) that have already been decided. In the second type, known as unsupervised learning, algorithms are taught to make predictions based on data provided to them without any guidance on the outcomes (target). Algorithms learn to generate predictions by establishing connections and patterns in previously unseen data throughout the training process. Similarities between the first two types of learning and the third form, Reinforcement Learning, end there. A goal-oriented agent investigates its surroundings. It makes a few choices as it navigates its environment. If Agent's choice advances Agent toward his goal, Agent will receive positive compensation; otherwise, Agent will receive negative compensation. Simply said, this is a method of discovery through experimentation [45]. Predictions regarding Chronic Diseases using categorization were a specific application of supervised learning in this study.

3.2. CLASSIFICATION TECHNIQUES

Classification is widely used since it is one of the most important methods of supervised machine learning. Classification can be broken down into two distinct categories. The first kind is utilized in the categorization of two groups, namely, the prognosis for, and is it Chronic Diseases? Therefore, the prediction is (Chronic Diseases or Not Chronic Diseases), and the form of classification that is based on just two classifications is referred to as Binary classification [48], [49].

3.2.1. Decision Tree Technique

It's an important supervised machine-learning algorithm for both classification and regression. Figure (3.2) depicts the DT as a tree-like flowchart in which the data is continuously partitioned along a particular axis. There are two parts to a decision tree: the nodes where features are tested and the papers that discuss the outcomes of those tests (the "root" node of the tree. Different decision tree methods (ID3, C4.5, C5, CART (Classification Regression), etc.) have different mathematical structures for segmenting training data[50, 51].

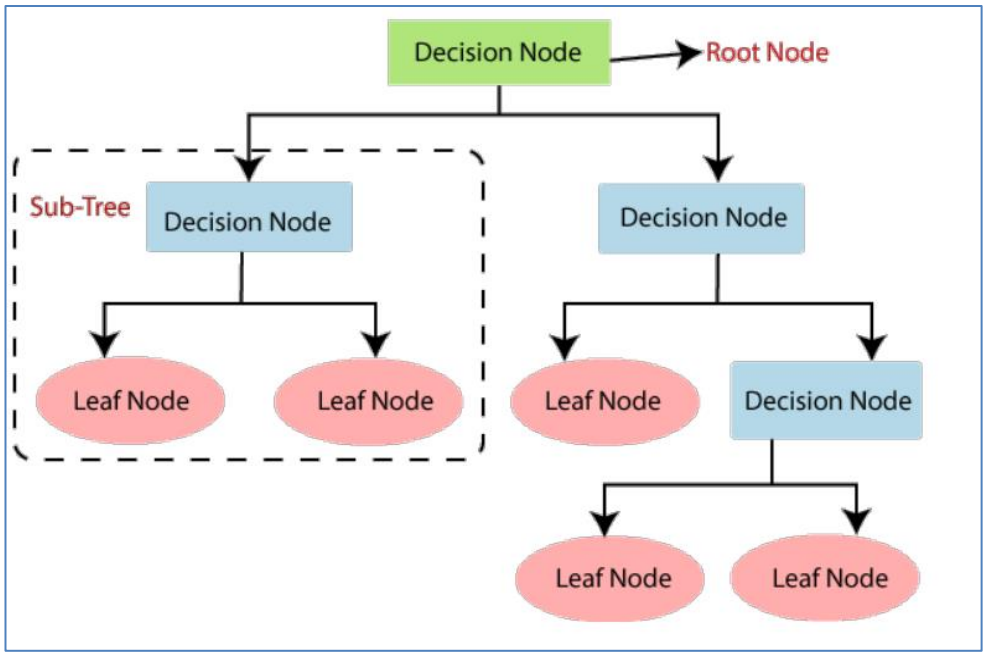

Figure 3.2. Methodological framework of DT [51].

One of these trees is generated by its algorithm depending on the type of reliable variable being used. A classification tree is constructed if the variable is categorical, whereas a regression tree is constructed if the variable is numeric. A categorical variable, Diseases state (Chronic Diseases or not Chronic Diseases) is the focus of the proposed model. This work thus produces categorization trees.

3.2.1.1. The DT Technique's Advantages

Uncomplicated and straightforward in its explanation. DT technique does not involve substantial data preparation. The DT method does not require an excessive amount of money to produce a tree. It is appropriate for use with both numerical and categorical sorts of data. It is effective when used with binary as well as many forecasts. One of the approaches known as the decision tree approach is categorized as a "white box" technique, which means that its operation may be simply comprehended. The effectiveness of the algorithm may be evaluated with the help of statistical measurements.

3.2.1.2. The Disadvantages of Using the DT Method

Overfitting is one of the most common issues that can arise while using the DT technique. Pruning, determining the bare minimum number of specimens that must be present in a leaf node, and measuring the depth of the tree are some of the tactics that are utilized in order to lessen the difficulty of this task. An unsteady decision tree is often the result of outliers. An effective method for dealing with this difficulty is the use of decision trees as part of an ensemble. Forecasts based on decision trees are piecewise constant approximations, not continuous or smooth estimates. XOR and equivalence problems are two examples of concepts that are difficult to state using DT. Biased trees can occur if the data set's categories are unevenly distributed.

3.2.2. Naïve Bayes Technique

I Bayes techniques are a family of supervised learning algorithms that take as their starting point Bayes' theorem and the "I" assumption of conditional independence

between any pair of features given the value of the class variable. I Bayesian models are helpful in medical research and text classification problems since they are easy to build and do not necessitate complex iterative parameter estimates. The I Bayesian classifier is commonly used despite its relative simplicity because it routinely outperforms other, more involved classification algorithms [52].

With the help of Bayes' theorem, one can figure out what the posterior probability is.P (j | i) from P(j), P(i), and P (I | j). The I Bayes classifier assumes that the weight assigned to each predictor (x) does not change the impact that x has on a particular class I. This presumption is known as "class conditional independence".

$$P(j|I,\ldots,i\_n)=(P(j)P(i\_1,\ldots,i\_n|j))/(P(i\_1,\ldots,i\_n)) \quad (3.1)$$

where P(I 1,...,I nj) is the probability of predictor given class, P(I 1,...,I n) is the prior probability of predictor, P(j) is the prior probability of class, and P(ji,...,I n) is the final probability of conditional probability [53, 54].

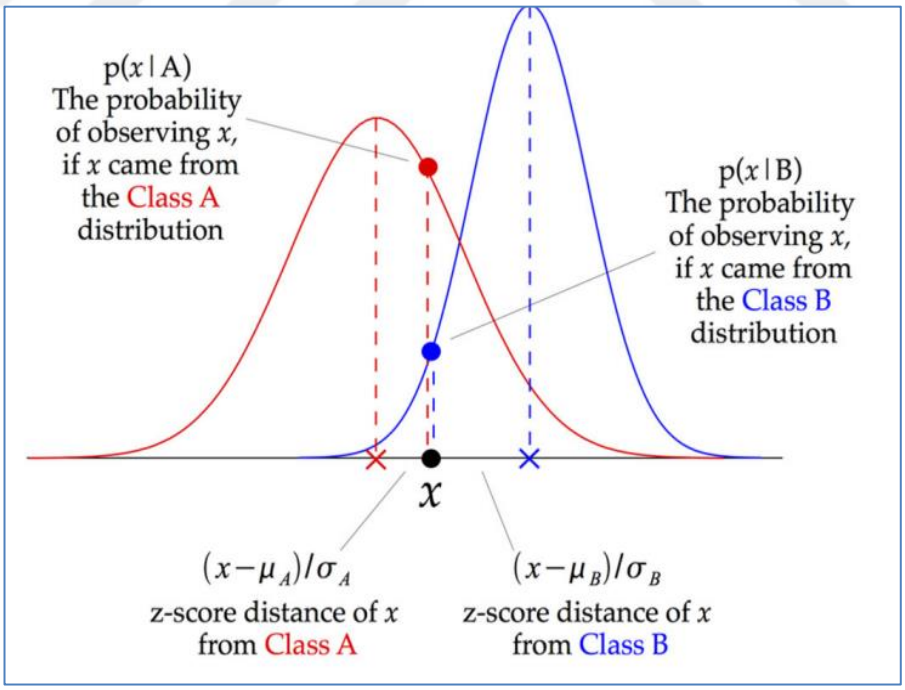

Figure 3.3. A classifier based on the Gaussian NB (55).

3.2.2.1. The NB Technique Advantages

It's not hard to figure out and implement. It can be put into action quickly. NB is an easy method that is quick and reliable. Simple to train with minimal data input. Very low price tag. It performs admirably when presented with a large dataset.

3.2.2.2. The Disadvantages of Using NB Technique

Independent predictors are quite difficult to obtain in practice. When no training tuples exist for a specific class, the posterior probability is at zero. The model's predictive abilities would be useless in this case. In mathematics, this problem is known as the Zero Probability/Frequency Issue.

3.2.3. Logistic Regression Technique

Supervised learning classification technique LR can assess the likelihood of an occurrence by fitting the data to a logistic curve, as depicted in Figure (3.4). A binary criterion to evaluate the result was employed. In logistic regression, a number of numerical or categorical predicted variables are utilized. Logistic regression is frequently utilized in healthcare and the social sciences. It’s also widely employed in the field of marketing, where it’s used to probe the buying tendencies of potential customers. Logistic regression is nothing more than a statistical log of how likely something is to occur. This function generates an S-shaped curve, which can be used for probability estimation of discrete values (0, 1, or yes/no) in relation to a given collection of explanatory factors [56],[57].

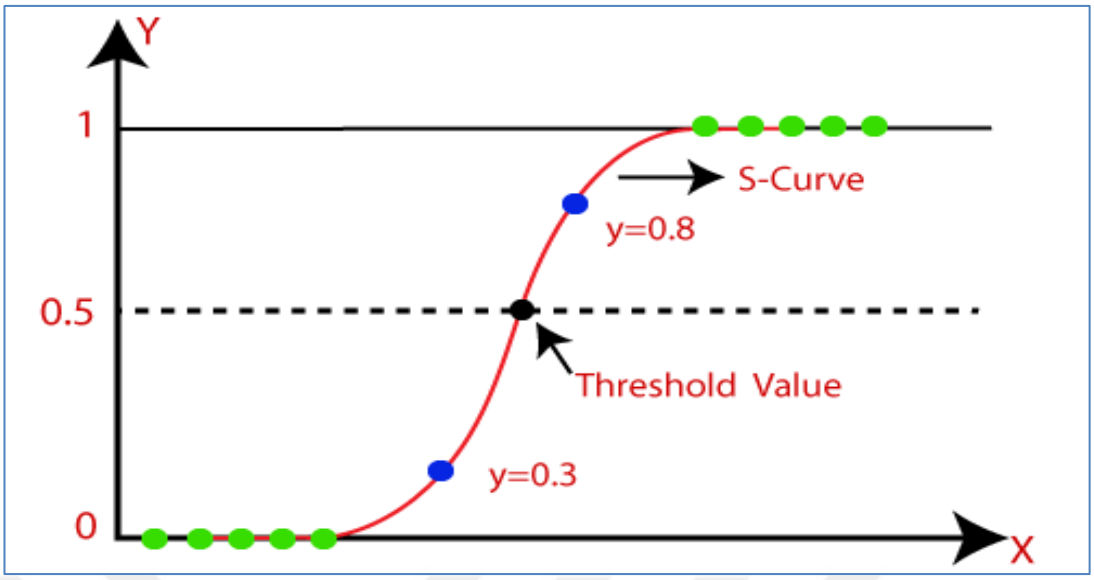

Figure 3.4. Method of LR's Logistic Curve (57).

Multivariate techniques include logistic regression. It seeks to establish a causal relationship between many independent predictor factors and a single dependent outcome measure. Two classes of membership outcomes (Anemic, Not-Anemic) were employed for prediction with binary LR. The main result of the LR model is the predicted odds or capability of a binary event, but the model may also offer additional information that must be taken into account. The probability of a "1" outcome is defined as being greater than 50% in a two-class situation. The number 0 represents all other possible categories. LR is a robust modeling technique, however it makes the simplifying assumption that the coefficients of interest between the response and predictor variables are linear. After gradually increasing the independent variables, coefficients were calculated [58]. The general form of the LR functional model for n independent variables is as follows:

$$p(X) = 1/1 + e^{-(B0+B1X1+B2X2+\cdots+BnXn)} = 1/1 + e^{-(b^t*X)} \quad (3.2)$$

Vector b, which for the considered data set correlates each record with the chance of diseases , was calculated utilizing a binary LR, where P(X) is the probability of diseases, X0, X1,, Xn are the predictor variables, B0, B1,, Bn are the regression coefficients.

3.2.3.1. The LR Technique Advantages

Its widespread use among data analysts and scientists can be attributed to the fact that it requires little in the way of computational resources, can be easily implemented and interpreted, and produces accurate results. And there's no need for scaling features either. In logistic regression, a probability score is assigned to each observation.

3.2.3.2. The Disadvantages of Using LR Technique

Unfortunately, it cannot process a wide variety of dimensional characteristics. This system is easily overfitted. Furthermore, non-linear features require a transformation because LR cannot address them directly. Logistic regression does not fare well with independent variables that are highly associated or comparable to one another but not to the target variable.

3.2.4. Stochastic Gradient Descent Learning

The use of this approach is an effective form of facilitation. Presented in simplified form, the SGD updates are as follows:

$$\theta^{((t+1))}=\theta^{((t))}- \alpha_t \nabla l_i (\theta^{((t))}) \quad (3.3)$$

Where:

Θ is the parameter updates
t is the iteration
α is the step size or learning size

In this case, the index I will change at each iteration. In practice, we frequently jumble the samples before going through them in order. Take note that the path can be easily extended to exploit the gradient of many samples., i.e. $\sum_{j=1}^{b} \nabla l_{i+j} (\theta^{(t)})$, Small-batch Gradient Descent is the proper term for this procedure. Mini-batching allows for quicker matrix operations and parallelization with more stable convergence[57].

When the volume of step becomes less pursuant to $\sum_t \alpha_t^2 < \infty \, and \, \sum_t \alpha_t = \infty \, , e.g., \alpha_t = \mathcal{O}(1/t)$, additionally, under mild conditions, SGD frequently converges to a regional minimum, and even a universal minimum, for a convex objective job. Linear convergence can be achieved by gradient descent and quadratic convergence by Newton's method under certain regularity conditions. This means that gradient descent must be used if the desired mission precision is to be $\mathcal{O}(log\,(1/\in))$ repetitions, and Newton's way occupies fewer. While, SGD occupies $\mathcal{O}\,(1/\in)$ repetitions, There's a good chance that's twice as bad as a gradient decline. Nonetheless, when n is sufficiently enough, assuming that the time complexity of computing the gradient of one specimen is a constant C, the overall time complexity of SGD equals is $\mathcal{O}\,(C/\in)$,that is weaker than that of gradient descent ,$\mathcal{O}(nC \log\,(1/\in))$ [59].

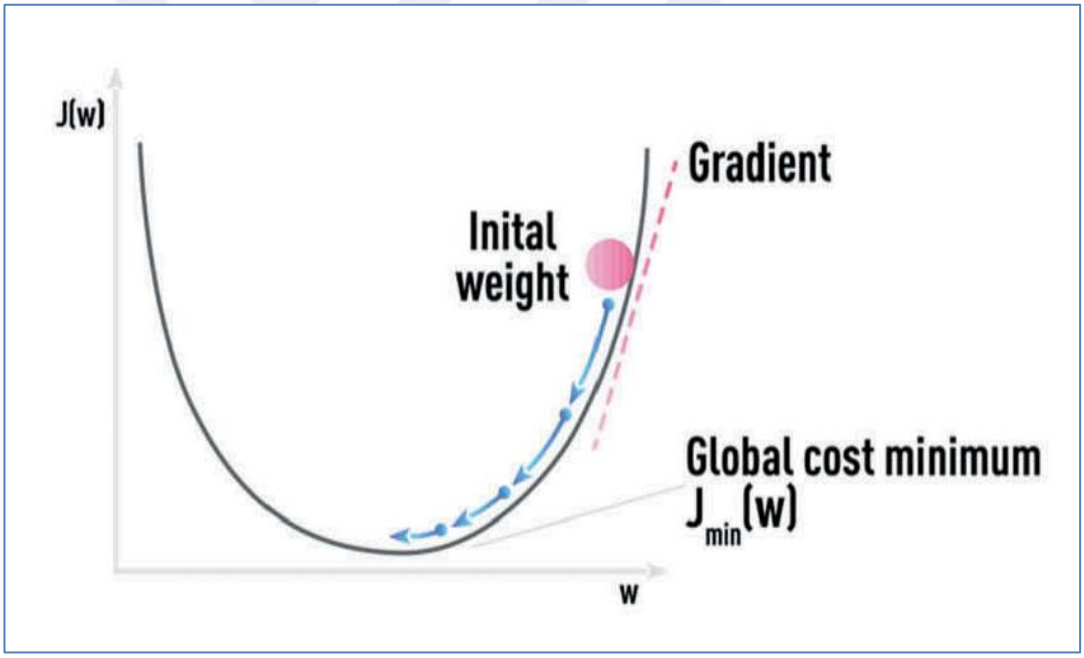

Figure 3.5. Schematic of stochastic gradient descent (SGD) (59).

3.2.4.1. The SGD learning Advantages

SGD has solved the large-scale, dispersed machine learning problems that arise most frequently in data categorization and NLP. Superior efficiency, Easy implementation and effective operation.

3.2.4.2. The Disadvantages of Using SGD Learning

The regularization and iteration frequency are examples of hyper parameters it wants. It retains sensitivity after a change in a defining feature.

3.2.5. K-Nearest-Neighbor (KNN) Approach

First proposed in 1951 by Evelyn Fix and Joseph Hodges, the non-parametric k-nearest Neighbor (KNN) method was further developed by Thomas Cover. In terms of complexity, it is among the lowest of supervised algorithms. Classification and regression analysis are two areas where it has been put to use. The k-nearest training examples from the dataset are used as input in both scenarios. When applied for either classification or regression, KNN yields distinct results. Membership in a class is the result of applying KNN classification. KNN regression output is the value of the thing in inquiry(51, 60). KNN requires two phases: Search through your training data for the K events that are most similar to the mystery event. Please choose the most appropriate labels for these K events.

Figure 3.6. Optimal KNN clustering (60).

Due to the scholastic nature of the results, the KNN approach was used as a classification methodology in this study (target). The KNN classifier is a nonparametric statistic since it does not presume anything about the probable distributions of the variables being employed. In Figure (3.6), the new item's neighbors are represented by the k symbol; the total number of nearby sounds is utilized to determine the object's (the target's) classification.

If k = 1, for instance, the nearest neighbor will be used to determine the class of the new object (target). Picking the right value for k can be done in numerous ways. In practice, though, it is usually easiest to simply try out a range of different values for k and pick the best one based on the KNN algorithm's performance. Until a new query is obtained, KNN delays making a generalization beyond the training instances, making it a lazy learning algorithm. The KNN algorithm can be used straight away without the need for a training set. Given an input and a k-value, KNN will utilize the training data set to classify the inputs. On display in Figure (3.6) are two groups, indicated by the blue circles and the red stars, respectively. You can think of these two sets as separate categories. These types of entities are reflected in the features space. The attribute has a two-dimensional representation; for example, the study's data might be thought of as belonging to one of two groups. The x and y coordinates are those. However, if there are three components, a three-dimensional environment is required. If we're working with an N-dimensional number, we'll need an N-dimensional space as well.

Adding a circle, representing fresh data, to the set of blue circles or red circles is called categorization. It was classified as a red star based on visual inspection of its closest neighbors, which revealed that it was, in fact, a red circle. This method of categorization is known as "nearest neighbor" since it uses this neighbor as a basis for the classification. Problems arise, however, since even though the red circles are the nearest, there may be an excess of blue circle clusters in the area. Therefore, the blue circles outnumber the red star and carry more weight at this place. Therefore, it is necessary to check whether or not the set contains the nearest k. Consider k = 3, which is closer to three samples. There being only one blue circle and two red circles, this means that the constellation is unambiguously a binary one. For this reason, it needs

to be reclassified as a red circle. Five blue circles and two red circles appear instead if k = 7. In this instance, it belongs in the same group as the blue circle. As a result, everything is subject to change depending on what k is [60], [61].

3.2.5.1. The KNN Technique Advantages

The method is straightforward to understand. It's a versatile tool that may serve both regression and classification needs. The degree of precision is quite great. There is no requirement to formulate any additional data assumptions, hone down on many parameters, or develop a model. When working with nonlinear data, it is extremely vital to remember this. The implementation of it does not require a significant amount of time.

3.2.5.2. The Disadvantages of Using the KNN Technique

The accuracy of the data plays a pivotal role. Depending on the size of the data set, this could take quite some time. Highly volume- and context-dependent. A large amount of storage space is required so that the entirety of the training set may be preserved. Since it keeps track of every training example, running it is a computationally intensive process.

3.3. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Networks (CNNs), a subtype of the discriminative deep architecture, are successful at analyzing two-dimensional data with a grid-like layout, such as photographs and movies. Deep 2D CNNs, which have numerous hidden layers and millions of parameters, may learn complicated objects and patterns given enough time and a big visual library annotated with ground-truth labels. Because of their special capabilities, they are the go-to instrument in many technical applications involving 2D signals like still photos and moving video frames, provided they receive the necessary training. The human visual cortex served as inspiration for the design of CNN. Numerous cells in the visual cortex have receptive fields that are tiny,

overlapping portions of the visual field. They filter the input space locally, with bigger receptive fields for more complicated cells[62[, [63].

3.3.1. The Architecture of CNN

The structure of the visual cortex in animals serves as inspiration for CNNs. The intermediate part of the network is composed of asynchronously connected convolutional layers (c-layers) and sub-sampling layers (s-layers), both of which are types of multi-layer neural networks. A c-layer design can be used to extract features when the input of each neuron is coupled to the local receptive field of the previous layer. Once all of the neighborhood traits have been retrieved, the relationship between them in space can be established. An s-layer is a special kind of layer that is only used to map such features. As a result of averaging the weights of the various feature mapping layers, we get a smooth surface. To reduce the overall size of the signal, we introduce the pooling approach, often known as sub-sampling or down-sampling [64]. Subsampling has been successfully utilized in audio compression to shrink file sizes without sacrificing sound quality. To further enhance the 2-D filter's position invariance, sub-sampling has also been implemented. In the max pooling method, a pooling function is utilized to replace the network output at a given position based on an aggregate of the statistics of the nearby outputs. Maximizing output in a rectangular area may be possible with the help of the max-pooling technique. When the input is translated, the pooled representation can remain unchanged. The addition of a max pooling layer between the convolutional ones allows for an increase in spatial abstractness that is proportional to the increase in feature abstractness. When using CNNs, filter training is all that's required as opposed to the time-consuming configuration required by traditional neural networks. Further, CNNs don't need any pre-training or human assistance to begin extracting characteristics immediately [65], [66]. Object recognition, computer vision, audio recognition, image classification, digitized handwriting recognition, face detection, behavior recognition, etc. are just a few of the many applications of CNNs.

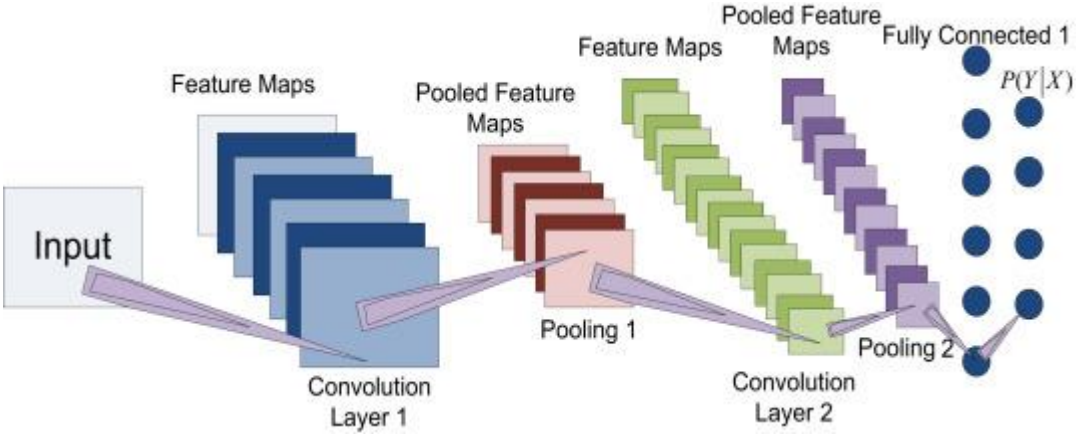

Figure 3.7. CNN's Straightforward Design Framework(64)

3.3.1. The Advantages of CNN

CNN has made significant strides in its development in recent years. In terms of recognition, CNN performs practically flawlessly, and this is due to the many benefits it offers. Convolutional neural networks (CNNs) are robust to shifts and distortions, allowing them to detect, for example, form changes brought on by factors other than the camera itself (pose, illumination, etc.). The CNN's invariant shift is possible thanks to its completely connected layers, which can be programmed to use the same weight in all directions. Since the same coefficients are used in multiple CNN layers, memory needs are low, in contrast to the traditional scenario where fully linked layers and the hidden layer both have features. Gains in Training Time and Accuracy Compared to More Conventional Neural Networks The normal neural network has a large number of parameters, which increases training time proportionally. Since the number of parameters in the CNN is much reduced, training time is also considerably cut down.

PART 4

METHODOLOGY

The proposed system utilizes both deep learning and conventional machine learning methods for the diagnosis and treatment of cardiovascular, renal, and diabetic diseases. In both cases, we first subject the data to some form of preprocessing, then use the two different classification techniques we've discussed, and lastly compare the outcomes. Figure (4.1) is a schematic depiction of the proposed system, which provides even more detail. Three different datasets were used to assess the proposed system (Pima Indians Diabetes Dataset, Cardiovascular Disease dataset, and UCI Heart Disease Data).

4.1. DATA PREPARATION

4.1.1. Platform Used

Python was used for all of this study's machine-learning procedures (3.83 version). The Pandas (version 1.0.5) libraries were used for data preprocessing, and the Scikit libraries were used to develop machine learning algorithms (version 0.23.1). The basic hardware configuration for this project is a PC with 24.00 GB of RAM and a 1.99 GHz Intel Core i7-8550U CPU running a 64-bit operating system on an x64-based processor.

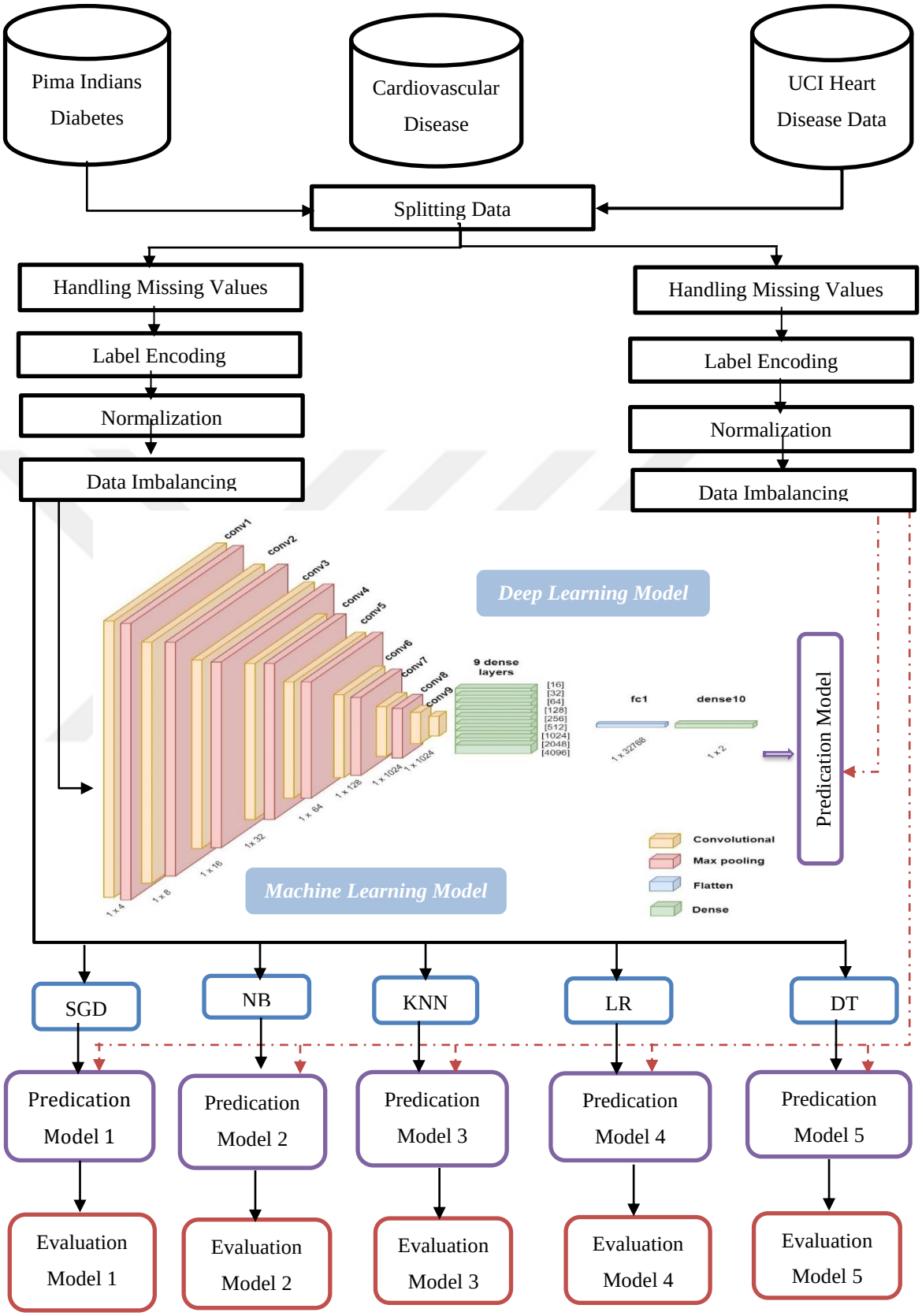

Figure 4.1. Flowchart representation of the model.

4.1.2. Data Collection

This investigation made use of three distinct datasets, which might be arranged in the following order:

4.1.2.1. Pima Indians Diabetes Database

Because of the kindness of the National Institute of Diabetes and Digestive and Kidney Diseases as well as Vincent Sigillito, a researcher at the Applied Physics Laboratory of the Johns Hopkins University, the diabetes dataset that pertains to the Pima aboriginals is now freely accessible to anyone interested in accessing it. The dataset was initially contributed by Sigillito, who is also credited as the donor. The author of this paper obtained genuine information by accessing the University of California, Irvine website on their own (University of California, Irvine). In the past, researchers have made use of these data to investigate potential vital signs that could be utilized to signal the presence of diabetes inside patients following the standards specified by the World Health Organization (WHO). The user will have the opportunity to practice on a total of 768 distinct training examples provided by this dataset. Eight features and a class variable are included in each training instance; these nine components, when put together, provide the training instance's label (see Table 4.1). Some of the characteristics that are taken into account are a diabetes pedigree function, triceps skin fold thickness in millimeters, diastolic blood pressure in millimeters of mercury, 2-hour serum insulin in milliunits per milliliter, body mass index in kilograms per meter squared, and years of age. The total number of live births, plasma glucose concentration, and total number of pregnancies are some other features. The value of the class variable can take on the values 0 or 1, with 0 denoting an individual who is healthy and 1 indicating a patient who is diabetic.

Table 4.1. Details on the Pima Indian Diabetes Dataset.

Feature	Description	Data type	Range
Preg	Number of times pregnant	Numeric	[0, 17]
Gluc	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (GTIT)	Numeric	[0, 199]
BP	Diastolic Blood Pressure (mm Hg)	Numeric	[0, 122]
Skin	Triceps skin fold thickness (mm)	Numeric	[0, 99]
Insulin	2-Hour Serum insulin (μh/ml)	Numeric	[0, 846]
BMI	Body mass index [weight in kg/(Height in m)]	Numeric	[0, 67.1]
DPF	Diabetes pedigree function	Numeric	[0.078, 2.42]
Age	Age (years)	Numeric	[21, 81]
Outcome	Binary value indicating non-diabetic /diabetic	Factor	[0,1]

4.1.2.2. Cardiovascular Disease Dataset

The dataset referring to cardiovascular sickness consists of 70,000 patient records, with the purpose (Cardio) of characterizing the presence or absence of heart disease through the employment of 11 factors that are detailed in Table. The dataset also pertains to cardiovascular disease (4.2). There are three distinct categories of input features: examination, objective, and subjective. Information that is factual falls under the category of objective features, whereas the results of a medical examination go under the category of examination features (containing information given by the patient). This dataset's cardio variable, which can be located in Table, is one that we are most interested in (4.2). There are a total of 70,000 records, 35,021 of which belong to people who have Cardio 0, and 34,979 of which belong to patients who have Cardio 1.

Table 4.2. Twelve attributes of the Cardiovascular Diseases dataset.

Attribute	Type	Description
Age	Continuous	**Age of the patient in days**
Gender	Discrete	**1: women, 2: men**
Height (cm)	Continuous	**Height of the patient in cm**
Weight (kg)	Continuous	**Weight of the patient in kg**
Ap_hi	Continuous	**Systolic blood pressure**
Ap_lo	Continuous	**Diastolic blood pressure**
Cholesterol	Discrete	**1: normal, 2: above normal, 3: well above normal**
Gluc	Discrete	**1: normal, 2: above normal, 3: well above normal**
Smoke	Discrete	**whether patient smokes or not**
Alco	Discrete	**Alcohol intake-Binary feature**
Active	Discrete	**Physical activity-Binary feature**
Cardio	**Discrete**	**Presence or absence of cardiovascular disease**

4.1.2.3. UCI Heart Disease Data

A multivariate dataset includes numerous independent mathematical or statistical variables, as well as multivariate numerical data analysis. This type of dataset can also be referred to as a multivariate database. The aforementioned criteria can both be satisfied by this particular dataset. The slope of the peak exercise ST segment, age, gender, the type of chest pain, the number of major vessels, Thalassemia, exercise-induced angina, old peak-ST depression brought on by exercise relative to rest, the slope of the peak exercise ST segment, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, and the highest heart rate that can be attained are some of the 14 characteristics. Even though this database contains 76 features, as shown in Table, only a subset of 14 of those characteristics have ever been the subject of an investigation in any study that has been published (4.3). The Cleveland database has been the sole resource that researchers working on ML have used in their work to this point. One of the most important objectives that this dataset seeks to achieve is the prediction of whether or not a patient suffers from cardiovascular disease. This prognosis is arrived at by taking into account the characteristics of the patient. Conducting a controlled experiment to study and uncover various insights that might be obtained from this dataset to possibly contribute to a

more thorough understanding of the situation is another technique that could be taken. This collection includes details on 920 unique patients in its entirety.

Table 4.3. Thirteen different characteristics of the Heart UCI dataset.

Attribute	Type	Description
Age	Continuous	**Age in years**
Cp	Discrete	**Chest pain type (4 values)**
Trestbps	Continuous	**Resting blood pressure (in mm Hg on admission to the hospital)**
Chol	Continuous	**Serum cholesterol in mg/dL**
Fbs	Discrete	**Fasting blood sugar > 120 mg/dL 1 = true; 0 = false**
Restecg	Discrete	**Resting electrocardiographic results (values 0,1,2)**
Thalach	Continuous	**Maximum heart rate achieved**
Exang	Discrete	**Exercise-induced angina (1 = yes; 0 = no)**
Oldpeak	Continuous	**ST depression induced by exercise relative to rest**
Slope	Discrete	**The slope of the peak Exercise ST segment (values 0,1,2)**
Ca	**Discrete**	**Number of major vessels (0–4) colored by fluoroscopy**

The three datasets mentioned above were used in two ways, the first was using the dataset as it is, and in the second case, the dataset was processed by using data augmentation.

4.2. DATA PRE-PROCESSING

The pre-processing of data is a stage that is necessary for both data mining and machine learning methodologies. Since data collected from the real world have a tendency to be inconsistent and noisy, and they may also lack data or contain material that is redundant or unnecessary. It has a detrimental impact on the efficiency of the algorithms and may lead to incorrect information and learning that is carried out improperly. During pre-processing, the data are cleaned up, scaled, and transformed into a format that is compatible with the algorithms that are being utilized, I and data balanced. Inaddition, there is a feature picker to choose the most beneficial characteristics. The stages of data preprocessing that were used in this investigation are illustrated in Figure (4.2).

4.2.1. Data Cleaning

At this point in the pre-processing processes, both the data that is missing and the data that is duplicated are examined. There are a few different approaches that can be taken to deal with the missing values, such as replacing the value with the attribute's median, mean, or mode. In this particular investigation, the average value of the feature column in question is substituted for any missing information concerning a particular feature. There were seven entries for residence, and there were ten values for short stature that were missing; this was adjusted by using the mean value of the feature column. In addition to this, it was made certain that there were no values that were identical to one another.

Figure 4.2. Data pre-processing phases.

4.2.2. Handel Missing Values

Researchers working in the field of machine learning run across the issue of missing values in datasets rather frequently. This is one of the most common challenges they face. Even when there is a large amount of data, the fraction of complete examples may be very low, and the majority of it may be missing values. This presents significant challenges for data mining and machine learning systems, which are

typically not designed to deal with missing values. The most common cause of mistakes and failures in learning systems is the absence of data, which also reduces the data's overall quality. As a consequence of this, addressing the issue of missing values is seen as an essential component of the overall data quality. As a result, the values that were lacking were considered to be the first step in the suggested process.

4.2.3. Encode Label

Frequently, datasets will have columns that contain string entries. On the other hand, fitting statistical models to such data frequently requires a numerical representation of all entries. This, in turn, involves the production of an encoding, also known as a vector representation of the entries. During the process of label encoding, the category value is changed to a numeric value that falls somewhere in the range of 0 and the number of classes minus 1.

4.2.4. Normalize Data

The process of transforming raw data values into another format that possesses characteristics that are better suitable for modeling and analysis is referred to as data normalization. The objective of normalization is to ensure that all genes are measured using the same unit of measurement for consistency's sake. As a direct consequence of this, it is utilized to prevent a disparity between the influence of small values and the effect of large values, which dominate the outcomes. There are a variety of techniques available for normalizing data, such as the min-max and z-score normalizations. The value is determined by applying the min-max normalization technique using equation (4.1).

$$\widehat{V} = \frac{V - min_a}{max_a - min_a} * (new\_max_a - new\_min_a) + new\_min_a \quad (4.1)$$

Where:

V: represents the gene value.

min_a : is the minimum original value.

max_a: is the maximum original value.

new_max_a and new_min_a: are the maximum and minimum interval of values.

4.2.5. Balancing Data

The dataset utilized was imbalanced, which led to bias to the majority classes; thus, to solve this problem, the Synthetic Minority Over-Sampling (SMOTE) technique [1] was utilized in order to overcome this challenge and get balanced data. The SMOTE technique was used in order to oversample the minority and balance the learning process and classes weights, which was very useful for the experimental results.

4.3. CLASSIFICATION ALGORITHMS

To train the recommended model, six different classification strategies were utilized. These strategies included the proposed ChronicCNN model, Stochastic Gradient Descent (SGD), Nave Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), and Logistic Regression (LR).

4.3.1. Chronic Disease Classification Based on the Proposed Chronic CNN Model

Convolutional neural networks (CNNs) are built with fully connected layers, pooling, and convolutional operations. A stack of numerous convolutional layers, a pooling layer, and one or more completely connected layers is a typical architectural design (see Table (4.4) for more information). The method of forward propagation is utilized at these stages to convert the input data into the output data. The ChronicCNN model under consideration includes each of the following layers:

- No. of Convolution with ReLU activation function: 9 layers
- No. of Max-pooling: 7 layers
- No. of Dense: 9 layers with ReLU activation function + 1 layer with softmax activation function
- No. of Flatten: 1 layer.

Table 4.3. ChronicCNN layers parameters.

Layer (type)	Output Shape	Param #
convld 1 (Conv1D)	(None, 8, 4)	8
max_poolingld1(MaxPooling1)	(None, 8, 4)	0
convld 2 (ConvID)	(None, 8, 8)	40
max_poolingld 2 (MaxPooling1)	(None, 8, 8)	0
convld 3 (Conv1D)	(None, 8, 16)	144
max poolingld 3 (MaxPooling1)	(None, 8, 16)	0
convld 4 (Conv1D)	(None, 8, 32)	544
max poolingld 4 (MaxPooling1)	(None, 8, 32)	0
convld 5 (Conv1D)	(None, 8, 64)	2112
max_poolingld 5 (MaxPooling1)	(None, 8, 64)	0
convld 6 (Conv1D)	(None, 8, 128)	8320
max_poolingld_6 (MaxPooling1)	(None, 8, 128)	0
convld 7 (Conv1D)	(None, 8, 1024)	132096
max poolingld 6 (MaxPooling1)	(None, 8, 1024)	0
convld 8 (Conv1D)	(None, 8, 1024)	1049600
convld 9 (Conv1D)	(None, 8, 125)	128125
dense_1 (Dense)	(None, 8, 16)	2016
dense_2 (Dense)	(None, 8, 32)	544
dense_3 (Dense)	(None, 8, 64)	2112
dense 4 (Dense)	(None, 8, 128)	8320
dense_5 (Dense)	(None, 8, 256)	33024
dense 6 (Dense)	(None, 8, 512)	131584
dense_7 (Dense)	(None, 8, 1024)	525312
dense_8 (Dense)	(None, 8, 2048)	2099200
dense 9 (Dense)	(None, 8, 4096)	8392704
flatten 1 (Flatten)	(None, 8, 32768)	0
dense_10 (Dense)	(None, 8, 2)	65538

4.3.2. Chronic Disease Classification Based on ML Classifiers

The dataset was divided into two parts before the ML approaches were implemented. The training set consisted of seventy percent of the data, and the test set comprised thirty percent of the total. After that, all of the ML methods that were investigated for this study were put into action to predict Chronic disease at different ages. This section provides an overview of five different machine-learning approaches that have been successfully applied to the diagnosis of chronic diseases, particularly the following:

Stochastic Gradient Descent (SGD): It is the method that is utilized for training data the majority of the time. Large networks and datasets result in longer training durations; hence, distributed SGD variations, most notably synchronous and asynchronous SGD, are often employed for training. The following is how this classifier is computed using equation (4.2):

$$e_{u.i} = r_{u.i} - \sum_{k=1}^{k} p_{uk} \cdot q_{ki} \quad (4.2)$$

Naïve Bayes (NB): This approach ranks among the top 10 in the field of data mining. This classifier places a strong emphasis on a probability metric while determining whether or not document A belongs in class B. It assumes that the presence or lack of one attribute has no influence on the presence or absence of another attribute, and it computes the results according to Eq. (4.3) as follows:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad (4.3)$$

Where;

$p(c|x)$: The posterior probability of the class (c, target) given a predictor (x, attributes).

$p(c)$: the historical probability of a particular class.

p$(x|c)$: the likelihood, or the probability that a predictor belongs to a specific class.

$p(x)$: the prior probability of the predictor.

K-Nearest Neighbor (KNN): It is one of the simplest and most straightforward ways of data mining. Learning and prediction analysis is carried out following the dataset or problem that is provided. The KNN classification model does not make any assumptions about the dataset; rather, it generates predictions only based on the values of the neighbor data. The "K" in KNN refers to the data values that are derived from the nearest neighbor. The KNN method analyzes a dataset to identify its "K" nearest neighbors, often known as its closest neighbors. The KNN model immediately classifies the dataset that was used for training. To locate an example that is comparable to Eq, the formula for Euclidean distance is utilized (4.4). The Euclidean

distance can be calculated by taking the square root of the sum of the squared differences that exist between the new instance (x-i.) and the existing instance (y-j.).

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{ik})} \quad (4.4)$$

Logistic Regression (LR): The logistic regression method, also known as the logistic model or the logit model, estimates the probability of an event taking place by conforming the available data to a logistic curve. It does so by examining the connection that exists between a category-dependent variable and several independent variables. According to Equations (4.5) and (4.6), the parametric model looks like this when B is a Boolean variable:

$$P(B = 1|A) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^{n} w_i A_i)} \quad (4.5)$$

$$P(B = 0|A) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i A_i)}{1+\exp(w_0 + \sum_{i=1}^{n} w_i A_i)} \quad (4.6)$$

Decision Tree (DT): This efficient approach can be utilized to successfully execute a wide range of tasks, some examples of which are pattern recognition, image processing, and machine learning. A set of fundamental tests are connected by the sequential model DT, which does this by successfully and fairly comparing a numerical attribute in each test to a threshold value. Because of their precision and the ease with which they can analyze data from a variety of sources, decision trees have a wide range of applications. The primary considerations for the decision tree are depicted in Equations (4.7) and (4.8):

$$Entropy = \sum_{i=1}^{c} p_i log p_i \quad (4.7)$$

Where C is the maximum number of nodes, and $\boldsymbol{p_i}$ is the probability of the node that is currently being considered.

$$Gain\ (S.A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \quad (4.8)$$

Where:

$\{S_1, \dots, S_i, \dots, S_n\}$ = partition of S according to the value of attribute A

n = number of attribute A

$|S_i|$ = number of cases in the partition Si

$|S|$ = total number of cases in S

4.4. PERFORMANCE MEASUREMENT

A confusion matrix was utilized to depict the performance of the various categorization strategies so that an evaluation of the effectiveness of the methods could be carried out. Metrics such as accuracy, precision, sensitivity, and specificity, as well as the F-score, are utilized during the review process.

4.4.1. Confusion Matrix

The effectiveness of the algorithm can be shown displayed through the usage of the confusion matrix. As can be seen in Figure (4.3), the confusion matrix for binary prediction problems is composed of two rows and two columns (class 0 and class 1), which classify the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (FN). A "1" was assigned to samples that were found to have Chronic disease, whereas a "0" was assigned to samples that did not have Chronic disease.

- True Positives (TP): Chronic diseases that are predicted as Chronic.
- True Negatives (TN): Not- Chronic diseases that are predicted as not- Chronic.
- False Positives (FP): Not- Chronic diseases that are predicted as Chronic.
- False Negatives (FN): Chronic diseases that are predicted as not- Chronic.

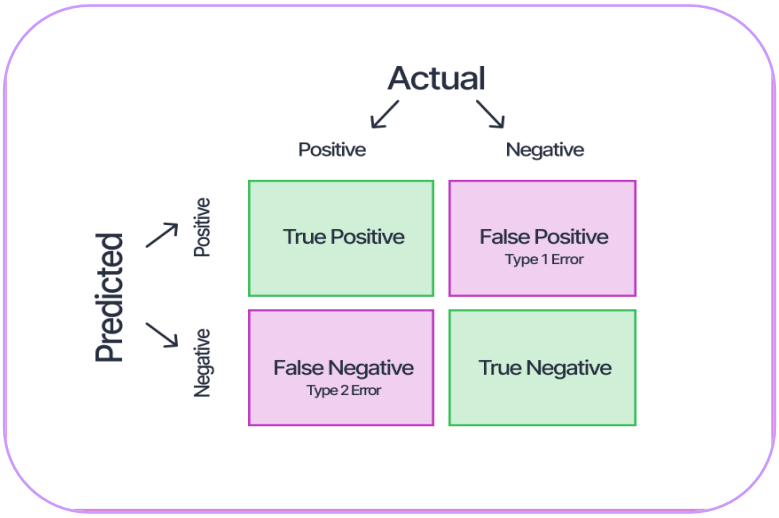

Figure 4.3. Confusion Matrix.

4.4.2. Accuracy

Classification accuracy is defined as the percentage of instances that the learner of the classification system properly categorizes (Chronic classified as Chronic and not-Chronic). is the proportion of samples that were successfully classified concerning the total number of samples that were examined. To calculate it, the equation that can be found below must be employed.

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP) \quad (4.9)$$

4.4.3. Precision

Precision or positive predictive value refers to the proportion of cases that are positive (Chronic classified as Chronic). The level of precision can be determined by dividing the number of accurate positive predicts (Chronic classified as Chronic) by the total number of positive predict (Chronic categorized as Chronic and not- Chronic) results. For the calculation, the equation shown below (4.10) is utilized.

$$\text{Precision} = TP/(TP + FP) \quad (4.10)$$

4.4.4. Recall

There are a few other names for it, including True Positive Rate (TPR), hit rate, and recall. It shows the percentage of correctly identified positive cases in comparison to the total number of positive instances. In the course of this inquiry, sensitivity was determined by applying the formula shown below, which is presented in Eq (4.11).

$$\text{Recall} = TP/(TP + FN) \tag{4.11}$$

4.4.5. F-Score

The score can range from 0 to 1, with 0 representing the lowest possible score and 1 representing the highest possible score. Therefore, the value 1 is used to represent F1, whereas the value 0 is used to represent F1's worst possible value. The following equations, (4.12) and (4.13), can be used to determine the F1 Score:

$$\text{F1 Score} = 2 * (Precision * Sensitivity/Precision + Sensitivity) \tag{4.12}$$

or

$$\text{F1 Score} = 2TP/(2TP + FP + FN) \tag{4.13}$$

PART 5

RESULTS AND DISCUSSION

5.1. EXPERIMENTS AND RESULTS

Using the procedures outlined in the figure, the experimental setup is constructed (4.1). The research team employed three different datasets, which they then split into training and testing sets. Then, the handling of the missing values, the encoding of the label, and the normalization of the data were done to rescale the data values of the dataset. In conclusion, a total of six different classification methods, including ChronicCNN as well as four or five different ML classifiers, are utilized to differentiate between healthy people and patients who have chronic diseases.

5.1.1. Results of The Proposed Chronic CNN Model

The following Table (5.1) displays the outcomes of applying this methodology to the aforementioned three datasets without augmentation and the with data augmentation:

Table 5.1. Results of the Chronic CNN model without data augmentation.

Dataset		Accuracy	Precision	Recall	F1-Score
Pima Indians Diabetes Dataset	Without augmentation	94	99.77	85.6	92.43
Cardiovascular Disease Dataset		88.01	99.74	58.54	73.85
UCI Heart Disease Dataset		79.3	99.78	45.32	62.37
Pima Indians Diabetes Dataset	With augmentation	99.81	99.08	99.82	99.8
Cardiovascular Disease Dataset		99.77	99.87	99. 76	99.75
UCI Heart Disease Dataset		99.92	99.94	99.93	99.9

In addition to the method of building the proposed model, which contributed to increasing the accuracy and reducing the testing time, the data pre-processing techniques had a significant impact in obtaining an ideal precision of 99.78%. This

was possible as a result of the significant impact the pre-processing techniques had on the data.

5.1.2. Results of ML Classifiers

The results of five ML methods on the three datasets are shown as follows:

5.1.2.1. Experimental Results on Dataset 1

It was determined that the following characteristics of Dataset 1 would be useful for making a diabetes prediction: Inclusion criteria are diabetes family history function, triceps skin fold thickness (mm), diastolic blood pressure (mmHg), 2-h serum insulin (mU/mL), body mass index (kg/m2), and age. Other characteristics include the overall number of pregnancies, the plasma glucose concentration, and the number of live births. The algorithms' results are tabulated below (5.2). SGD and LR achieved 94% accuracy, the best of any methods tested. With an accuracy of 84.000 percent, RF ranks second.

Table 5.2. A comparison of the various ML algorithms' performance on Dataset 1.

Classifier		Accuracy	Precision	Recall	F1-Score
Stochastic Gradient Descent (SGD)	Without Augmentation	94	94	70	79
Naïve Bayes (NB)		79	79	68	72
Random Forest (RF)		84	83	84	83
K-Nearest Neighbor (KNN)		69	69	68	68
Logistic Regression (LR)		94	94	70	79
Decision Tree (DT)		58	58	59	58
Stochastic Gradient Descent (SGD)	With Augmentation	74	77	73	75
Naïve Bayes (NB)		77	75	79	77
Random Forest (RF)		81	79	82	81
K-Nearest Neighbor (KNN)		77	76	78	72
Logistic Regression (LR)		75	71	76	75
Decision Tree (DT)		77	79	77	78

5.1.2.2. Experimental Results on Dataset 2

In this part, Cardio was predicted by employing features that comprise factual information, whereas examination features include the findings of a medical exam with 70,000 records in total, 35,021 of which belong to persons who have Cardio 0 and 34,979 to patients who have Cardio 1. The RF algorithm accomplished the maximum level of accuracy, achieving 84.00 percent. The findings are laid forth in the table below (5.3).

Table 5.3. A comparison of the various ML algorithms' performance on Dataset 2.

Classifier		Accuracy	Precision	Recall	F1-Score
Stochastic Gradient Descent (SGD)	Without Augmentation	65	64	67	63
Naïve Bayes (NB)		79	79	80	79
Random Forest (RF)		84	83	84	83
K-Nearest Neighbor (KNN)		70	70	70	70
Logistic Regression (LR)		62	60	74	54
Decision Tree (DT)		77	77	77	77
Stochastic Gradient Descent (SGD)	With Augmentation	88	89	87	88
Naïve Bayes (NB)		64	77	61	69
Random Forest (RF)		93	95	91	93
K-Nearest Neighbor (KNN)		88	90	87	88
Logistic Regression (LR)		63	78	60	68
Decision Tree (DT)		87	90	88	88

5.1.2.3. Experimental Results on Dataset 3

To the slope of the peak exercise ST segment, add the following 14 factors: age, sex, chest pain type, number of major vessels, Thalassemia, exercise-induced angina, old peak-ST depression brought on by exercise relative to rest, the slope of the peak exercise ST segment, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, and maximum heart rate. Even though there are 76 qualities in total, just 14 of them are examined here.

It can be shown in Table (5.4) that DT, KNN, RF, NB, and LR have all experienced significant progress since the beginning of the study. Using feature selection methods, LR achieved the greatest result with a 70.00 percent accuracy rate, while KNN and NB may be considered the best among the rest of the classifiers as they yielded a 68 percent accuracy rate, and the SGD classifier was the worst in accuracy. Both KNN and NB can be considered the best among the rest of the classifiers because they yielded an accuracy rate of 68 percent.

Table 5.4. A comparison of the various ML algorithms' performance on Dataset 3.

Classifier		Accuracy	Precision	Recall	F1-Score
Stochastic Gradient Descent (SGD)	Without Augmentation	56	61	61	56
Naïve Bayes (NB)		68	56	61	56
Random Forest (RF)		64	57	57	57
K-Nearest Neighbor (KNN)		68	62	63	62
Logistic Regression (LR)		70	54	70	49
Decision Tree (DT)		60	57	56	56
Stochastic Gradient Descent (SGD)	With Augmentation	74	77	73	75
Naïve Bayes (NB)		77	75	79	77
Random Forest (RF)		81	79	82	81
K-Nearest Neighbor (KNN)		77	76	78	77
Logistic Regression (LR)		75	71	78	75
Decision Tree (DT)		77	79	77	78

5.2. COMPARISON BETWEEN DL AND ML RESULTS

Because the structure of the suggested model was quite correct, as was stated before, it should come as no surprise that the outcomes of the first way are superior to those of the second. The figure displays the comparison chart that was requested (5.1).

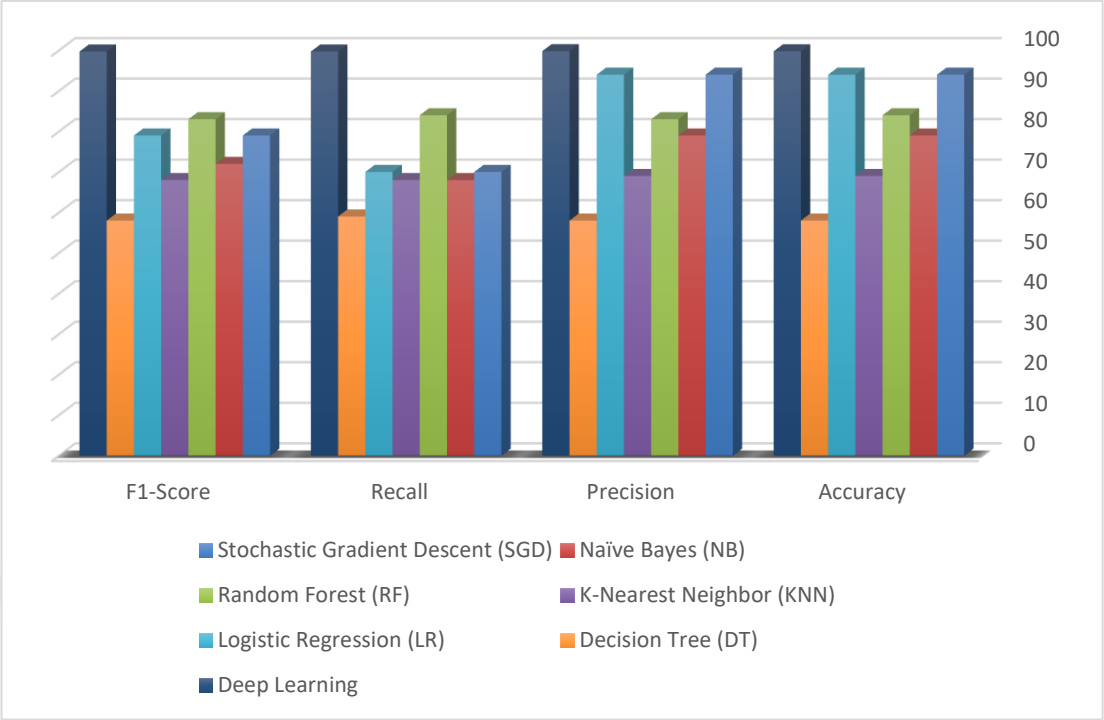

Figure 5.1. Comparing the outcomes of DL and ML on dataset 1.

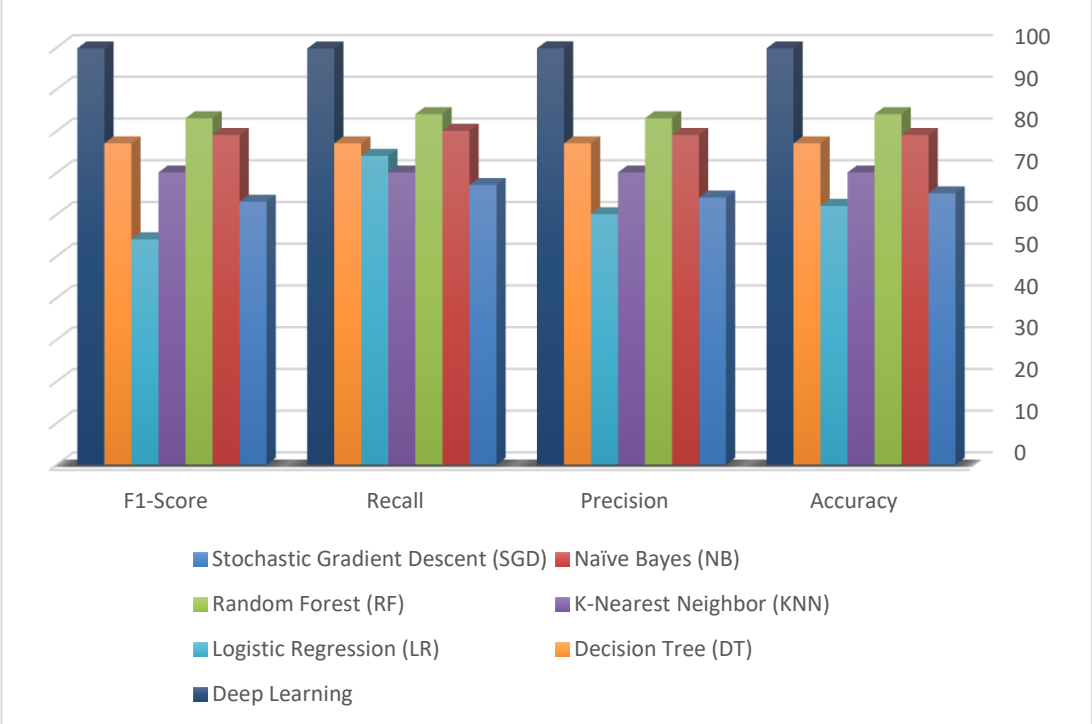

Figure 5.2. Comparing the outcomes of DL and ML on dataset 2.

When compared to the other six ML classifiers, the results presented in the preceding Figure (5.2) make it abundantly clear that the deep learning method, which is

predicated on a one-dimensional convolutional neural network, provided the highest level of accuracy in the detection of chronic disease.

The results that are presented in the following figure (5.3) make it abundantly clear that the deep learning method, which is predicated on a one-dimensional convolutional neural network, yielded the best chronic disease detection accuracy when compared to the other six ML classifiers. This is the case when the results are compared to one another.

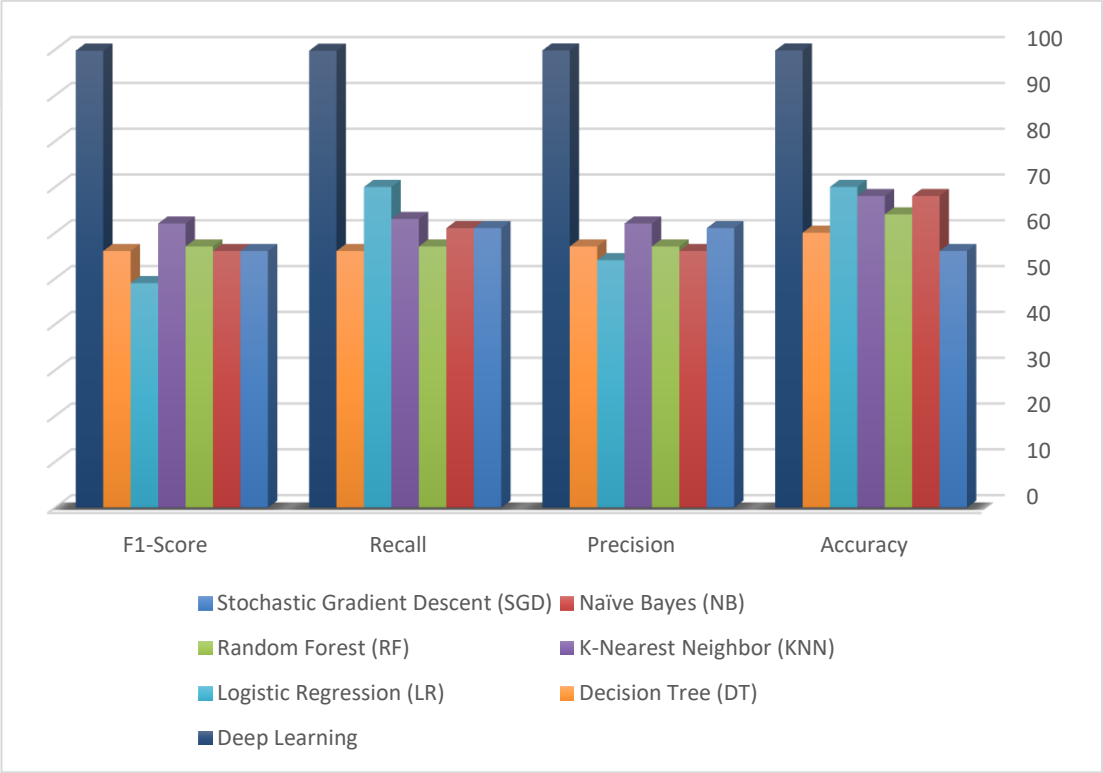

Figure 5.3. Comparing the outcomes of DL and ML on dataset 3.

5.3. COMPARISON BETWEEN THE PROPOSED CHRONICCNN MODEL WITH OTHER STUDIES

On Dataset 1, the performance of the proposed chronicCNN strategy was analyzed and compared based on accuracy with many research; the findings revealed that the suggested chronicCNN is the most effective method without the requirement of feature selection procedures, as shown in Table (5.5).

Table 5.5. Accuracy of the proposed ChronicCNN model with other studies when implemented on data set 1.

Ref. No.	Year	Technique	Dataset	Accuracy	Precision
Madan et al. (11)	2022	CNN-Bi-LSTM	PIMA Indians Diabetes	98%	87%
Proposed ChronicCNN	2022	CNN	PIMA Indians Diabetes	94% 99.81	99.77% 99.08

The best results may be seen when using the suggested ChronicCNN on Dataset 2 as shown in Table (5.6). Yet the RF method has strong rivals in the industry. The proposed ChronicCNN has the highest accuracy among the several classification methods used in the study. The proposed ChronicCNN model presented also achieves the maximum accuracy across all datasets.

Table 5.6. Accuracy of the proposed ChronicCNN model with other studies when implemented on dataset 2, 3.

Ref. No.	Year	Technique		Dataset	Accuracy	Precision
Mehmood et al., (29)	2021	CNN		UCI Heart Disease Dataset	97%	/
Proposed ChronicCNN	2022	CNN	Without Augmentation	UCI Heart Disease Dataset	79.3%	99.78%
				Cardiovascular Disease Dataset	88.01%	99.74%
Proposed ChronicCNN	2022	CNN	With Augmentation	UCI Heart Disease Dataset	99.92%	99.94%
				Cardiovascular Disease Dataset	99.77%	99.87%

5.4. DISCUSSION

In the diagnosis of chronic disease and the exploration of the knowledge of social aspects connected with it, the techniques of machine learning have produced some promising results. The results of this study demonstrated that socio-demographic characteristics are capable of serving as reliable indicators of chronic disease. One can get the following conclusion: there is a direct correlation between the various aspects that have affected the findings.

The performance of the six different classification strategies was evaluated and compared based on accuracy; the results showed that DL is the best way without the need for feature selection techniques, as displayed in Figure (5.4). Figure 5.5 demonstrates that when applying feature selection strategies to Dataset 1, the DL method yields the greatest results. The RF and NB techniques, however, are extremely close competitors. When we compare the precision of the various classification methods employed in the research, we find that DL has the highest precision, while LR has the best second precision. These findings are presented in Figure (5.6). In addition, the deep learning model shown has the highest accuracy across all of the datasets.

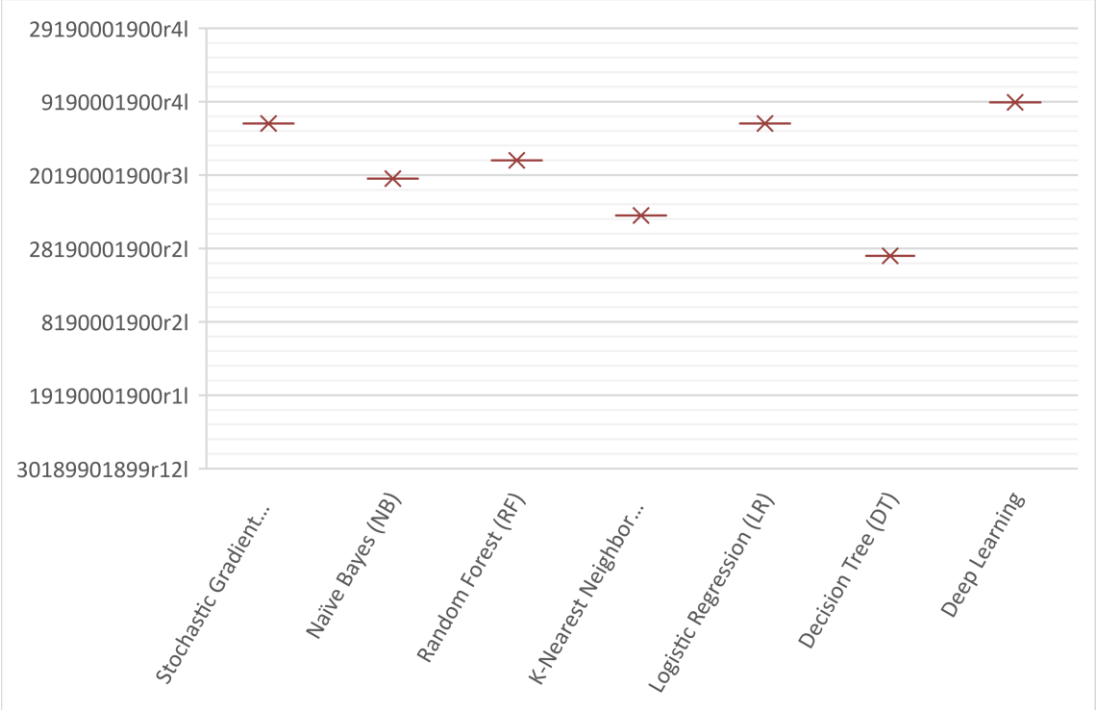

Figure 5.4. Accuracy of DL and ML techniques when implemented on data set 1.

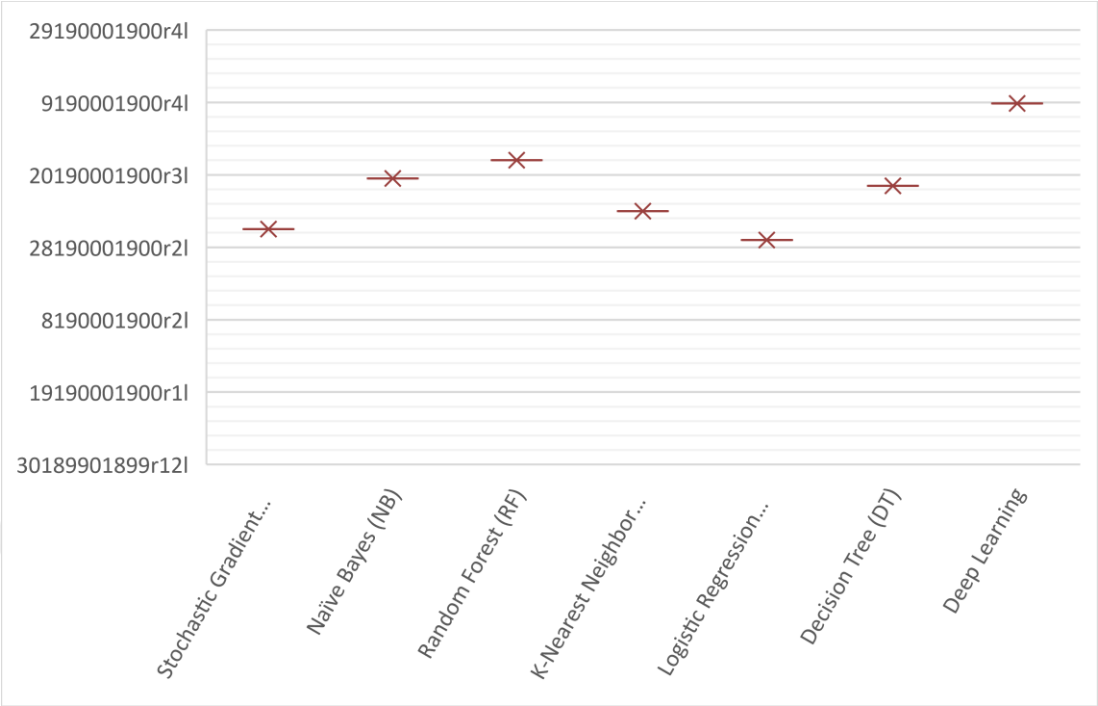

Figure 5.5. Accuracy of DL and ML techniques when implemented on data set 2.

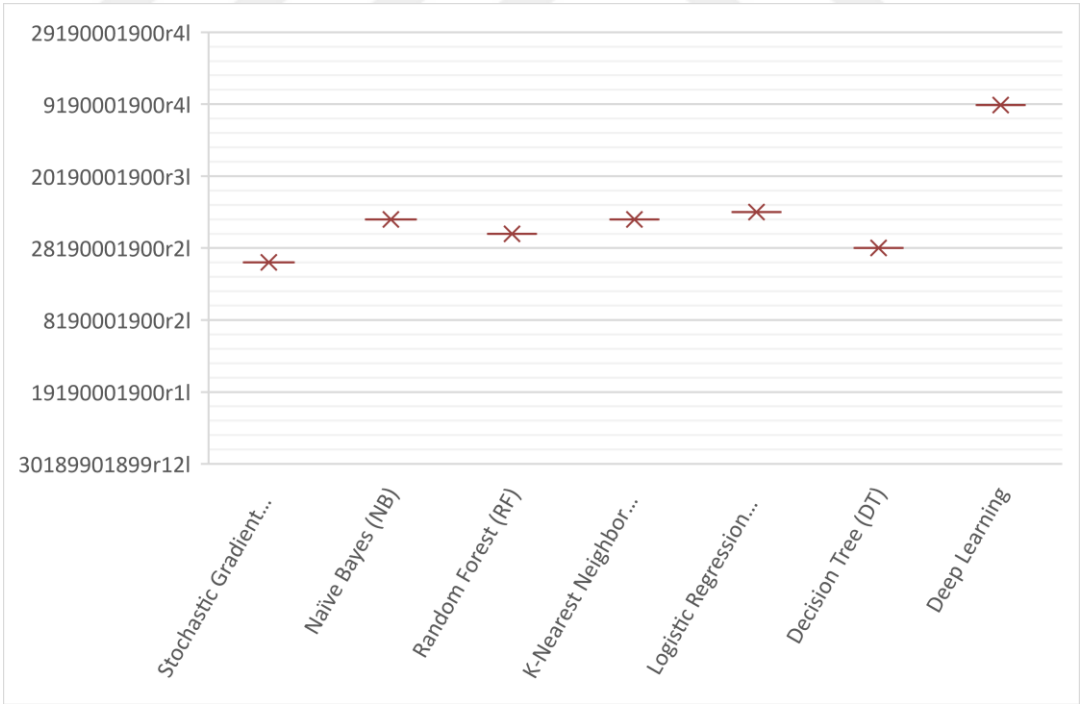

Figure 5.6. Accuracy of DL and ML techniques when implemented on data set 3.

The charts of accuracy and loss of the proposed system will be shown in Fig. (5.7) and (5.8) for data before the data augmentation process.

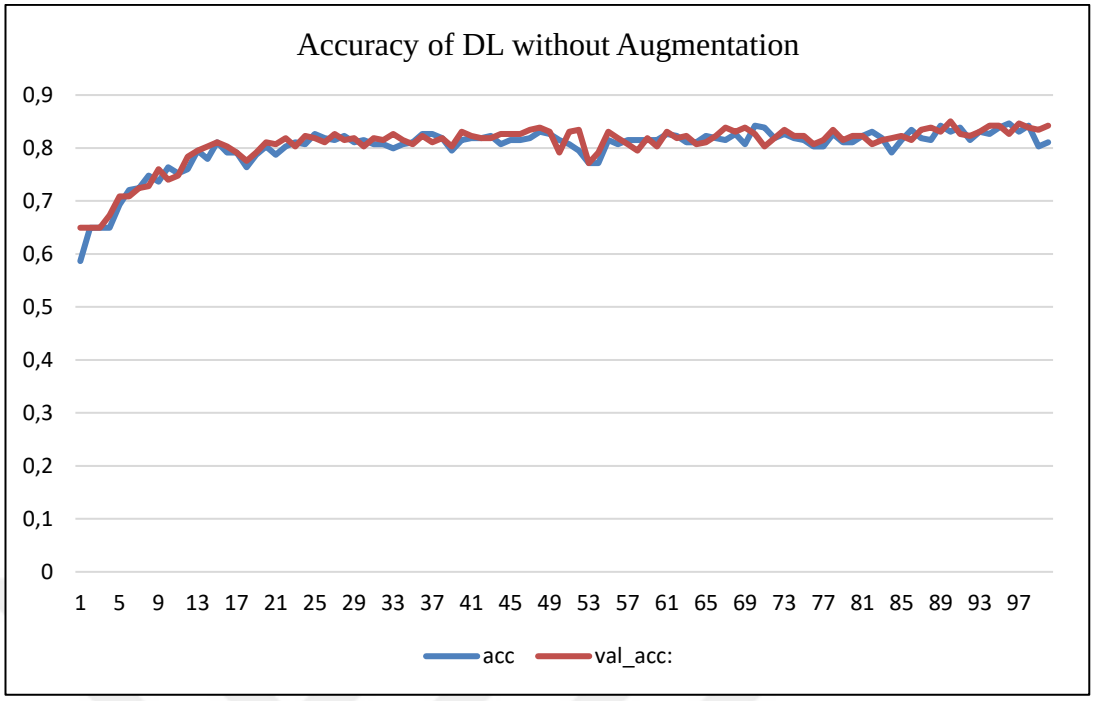

Figure 5.7. Accuracy of the proposed DL without augmentation.

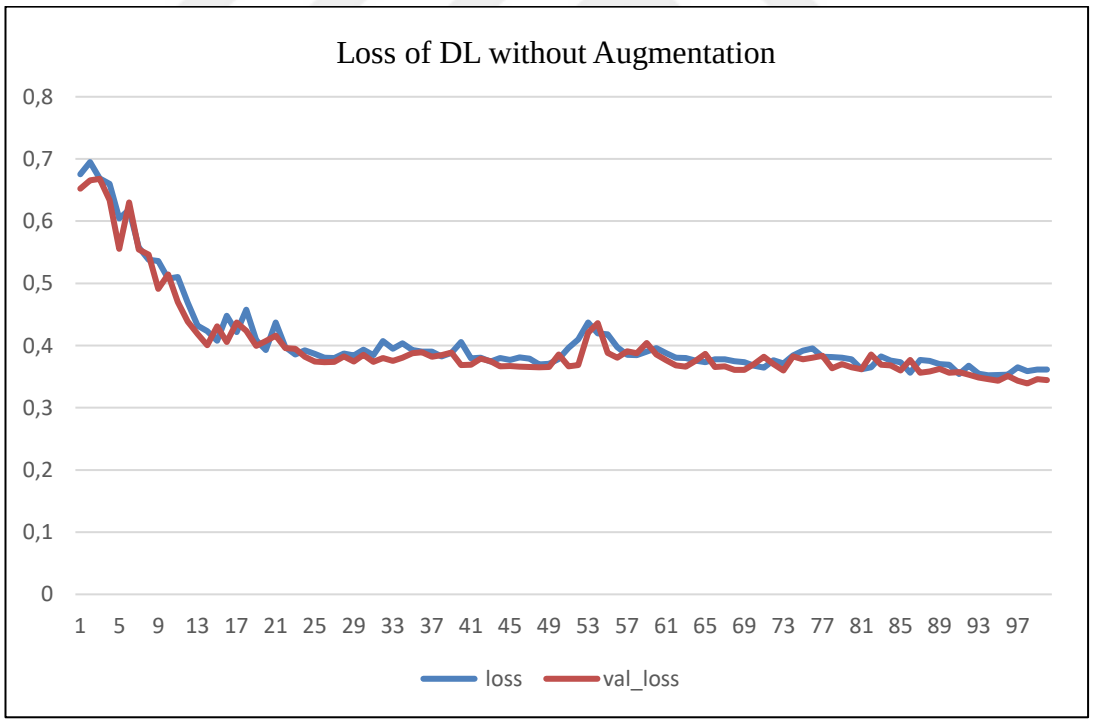

Figure 5.8. Loss of the proposed DL without augmentation.

The charts of accuracy and loss of the proposed system will be shown in Fig. (5.9) and (5.10) for data before the data augmentation process.

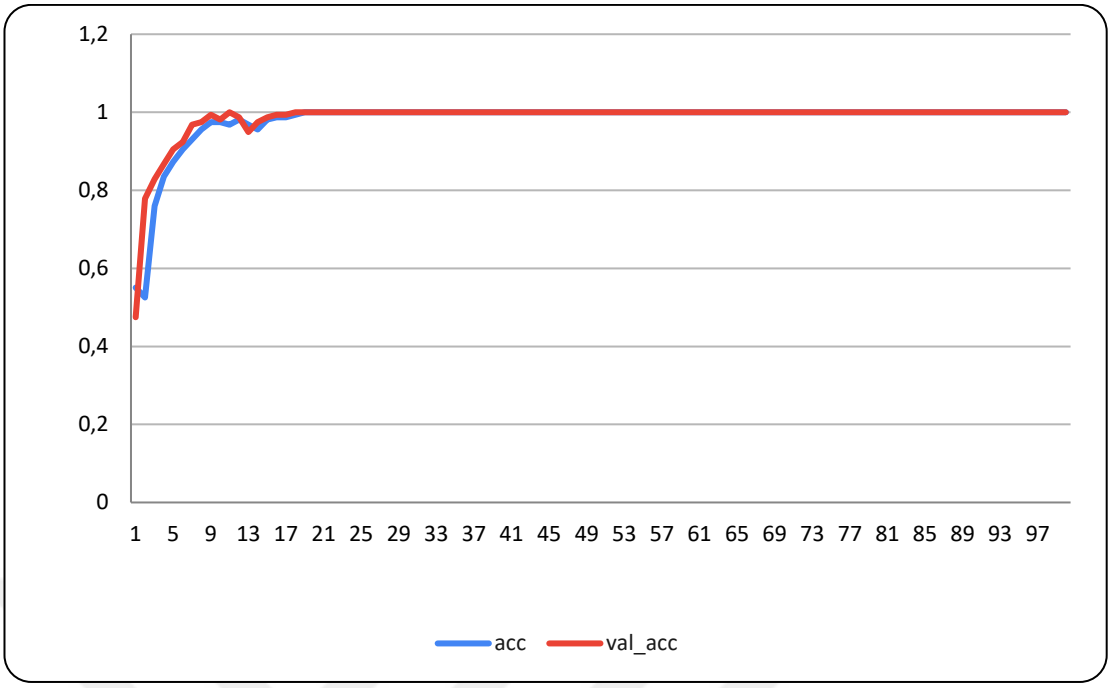

Figure 5.9. Accuracy of the proposed DL with augmentation.

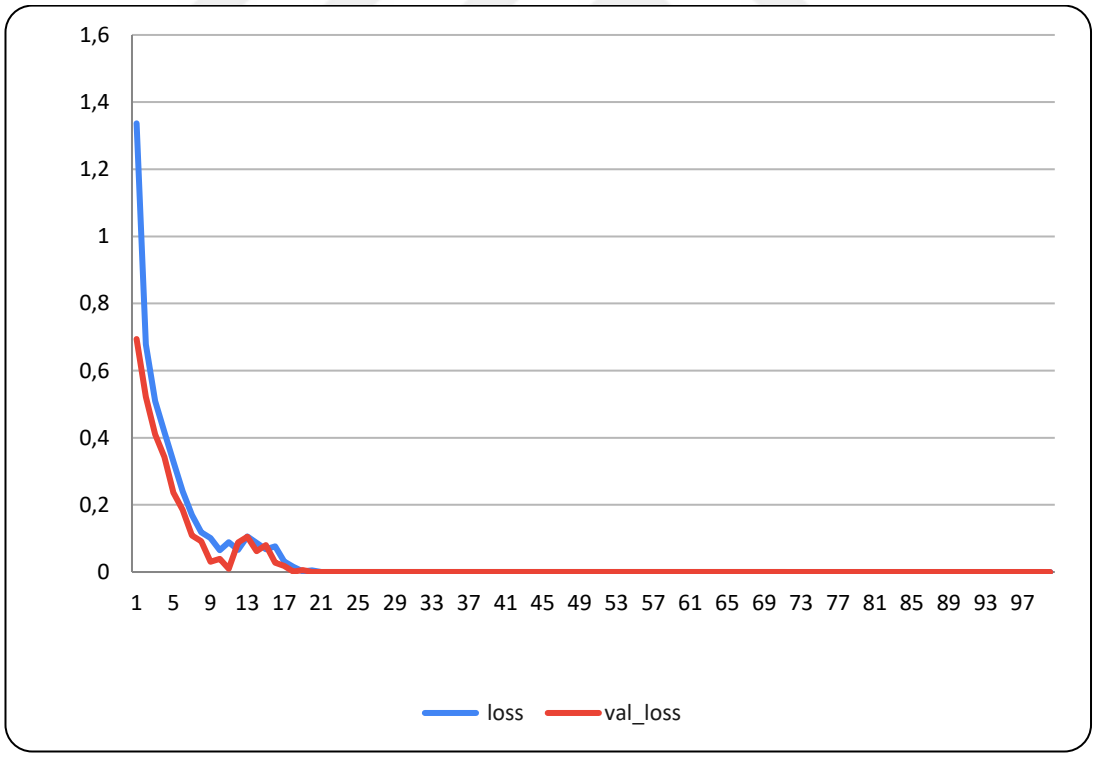

Figure 5.10. Loss of the proposed DL with augmentation.

PART 6

CONCLUSION

The high frequency of older individuals who have several chronic diseases makes it extremely difficult for public health professionals to maintain older people's health in the community. According to this study, deep learning and machine learning could aid in the detection of chronic diseases. This system is divided into various steps, including the stages of disease classification and data pre-processing. One method uses a deep convolution neural network (CNN), and the other uses five machine learning algorithms: stochastic gradient descent (SGD), Nave Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). The pre-processing stage of data is also a part of this system. The proposed model uses three data sets to categorize kidney, heart, and diabetes problems, including the Pima Indians Diabetes Dataset, the Cardiovascular Disease Dataset, and the UCI Heart Disease Data. The proposed ChronicCNN model yielded the best results; without employing data augmentation, accuracy for datasets 1, 2, and 3 was 94%, 79.3%, and 88.01%, while precision was 99.77%, 99.78%, and 99.74%. The accuracy results with data augmentation were (99.81%, 99.92%, and 99.77%) and the precision results were (99.08%, 99.94%, and 99.87%) for datasets 1, 2, and 3, respectively. While the best classifier for both SGD and LR classifiers was found to have a precision of 61%.

Future work will quantify the variance of the model outputs and use a variety of datasets to test the accuracy of our proposed models. Additionally, we will apply methods like feature selection and k-fold cross-validation to the models that have been given. The preprocessing and feature selection procedures used in the suggested model allowed for a fairly quick implementation of the proposed system on a computer device with limited capabilities.

REFERENCES

1. SIBONI, Fatemeh Samiei, et al. Quality of life in different chronic diseases and its related factors. International journal of preventive medicine, 2019, 10.

2. Kelly, Bridget B., Jagat Narula, and Valentín Fuster. "Recognizing global burden of cardiovascular disease and related chronic diseases." *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine* 79.6 (2012): 632-640.

3. Grover A, Joshi A. An overview of chronic disease models: a systematic literature review. Vol. 7, Global journal of health science. 2015.

4. Schwinger RHG. Pathophysiology of heart failure. Vol. 11, Cardiovascular Diagnosis and Therapy. 2021.

5. Banday MZ, Sameer AS, Nissar S. Pathophysiology of diabetes: An overview. Avicenna J Med. 2020;10(04).

6. Ammirati AL. Chronic kidney disease. Vol. 66, Revista da Associacao Medica Brasileira. Associacao Medica Brasileira; 2020. p. 3–9.

7. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. Vol. 398, The Lancet. Elsevier B.V.; 2021. p. 786–802.

8. Ling Y, An Y, Liu M, Hu X. An error detecting and tagging framework for reducing data entry errors in electronic medical records (EMR) system. In: Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013. 2013.

9. Gulmez H. Detection of chronic disease in Primary Care Using Artificial Intelligence Techniques. In: Computational Intelligence and Soft Computing Applications in Healthcare Management Science. 2020.

10. Nithyalakshmi V, Sivakumar DrR, Sivaramakrishnan DrA. Automatic Detection and Classification of Diabetes Using Artificial Intelligence. International Academic Journal of Innovative Research. 2021 Dec 20;8(1):01–5.

11. Chaki J, Thillai Ganesh S, Cidham SK, Ananda Theertan S. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Vol. 34, Journal of King Saud University - Computer and Information Sciences. 2022.

12. Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4).

13. Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, et al. An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. Applied Sciences [Internet]. 2022;12(8). Available from: https://www.mdpi.com/2076-3417/12/8/3989

14. Amisha Singla, Ankit Kumar, Surbhi Tyagi, Garima Gupta. DIABETES PREDICTION MODEL. International Research Journal of Modernization in Engineering Technology and Science. 2022 May;4(5).

15. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl. 2022;

16. Barik S, Mohanty S, Mohanty S, Singh D. Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: Smart Innovation, Systems and Technologies. Springer Science and Business Media Deutschland GmbH; 2021. p. 399–409.

17. Spoorthy Y, Sunitha T. Diabetes Prediction in Women using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT). 2021;

18. Alaa Khaleel F, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. Mater Today Proc. 2021;

19. Zhou H, Myrzashova R, Zheng R. Diabetes prediction model based on an enhanced deep neural network. EURASIP J Wirel Commun Netw. 2020;2020(1).

20. Challa M, Chinnaiyan R. Optimized machine learning approach for the prediction of diabetes-mellitus. In: Advances in Intelligent Systems and Computing. 2020.

21. Choudhury A, Gupta D. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In: Advances in Intelligent Systems and Computing. 2019.

22. Aminah R, Saputro AH. Diabetes prediction system based on iridology using machine learning. In: 2019 6th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2019. 2019.

23. Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. In: 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings. 2019.

24. Swapna G, Soman KP, Vinayakumar R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. In: Procedia Computer Science. 2018.

25. Morris SA, Lopez KN. Deep learning for detecting congenital heart disease in the fetus. Nat Med. 2021;27(5).

26. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M, et al. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. Applied Intelligence. 2019 Jan 15;49(1):16–27.

27. Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular Disease Detection using Ensemble Learning. Javed AR, editor. Comput Intell Neurosci [Internet]. 2022;2022:5267498. Available from: https://doi.org/10.1155/2022/5267498

28. Sharifrazi D, Alizadehsani R, Joloudari JH, Band SS, Hussain S, Sani ZA, et al. CNN-KCL: Automatic myocarditis diagnosis using convolutional neural network combined with k-means clustering. Mathematical Biosciences and Engineering. 2022;19(3):2381–402.

29. Hussain S, Nanda SK, Barigidad S, Akhtar S, Suaib M, Ray NK. Novel Deep Learning Architecture for Predicting Heart Disease using CNN. In: Proceedings - 2021 19th OITS International Conference on Information Technology, OCIT 2021. Institute of Electrical and Electronics Engineers Inc.; 2021. p. 353–7.

30. Mehmood A, Iqbal M, Mehmood Z, Irtaza A, Nawaz M, Nazir T, et al. Prediction of Heart Disease Using Deep Convolutional Neural Networks. Arab J Sci Eng. 2021;46(4).

31. Zhang X, Gu K, Miao S, Zhang X, Yin Y, Wan C, et al. Automated detection of cardiovascular disease by electrocardiogram signal analysis: A deep learning system. Cardiovasc Diagn Ther. 2020;10(2).

32. Shankar VV, Kumar V, Devagade U, Karanth V, Rohitaksha K. Heart Disease Prediction Using CNN Algorithm. SN Comput Sci. 2020;1(3).

33. Shubhanshi Singhal, Harish Kumar, Vishal Passricha. Prediction of Heart Disease using CNN. American International Journal of Research in Science, Technology, Engineering & Mathematics. 2018;23(1).

34. Wang YN, Ma SX, Chen YY, Chen L, Liu BL, Liu QQ, et al. Chronic kidney disease: Biomarker diagnosis to therapeutic targets. Vol. 499, Clinica Chimica Acta. 2019.

35. Mondol C, Shamrat FMJM, Hasan MR, Alam S, Ghosh P, Tasnim Z, et al. Early Prediction of Chronic Kidney Disease: A Comprehensive Performance Analysis of Deep Learning Models. Algorithms. 2022 Sep 1;15(9).

36. Al-Momani R, Al-Mustafa G, Zeidan R, Alquran H, Mustafa WA, Alkhayyat A. Chronic Kidney Disease Detection Using Machine Learning Technique. In: 2022 5th International Conference on Engineering Technology and its Applications (IICETA). 2022. p. 153–8.

37. Ilyas H, Ali S, Ponum M, Hasan O, Mahmood MT, Iftikhar M, et al. Chronic kidney disease diagnosis using decision tree algorithms. BMC Nephrol. 2021 Dec 1;22(1).

38. Elkholy SMM, Rezk A, Saleh AAEF. Early Prediction of Chronic Kidney Disease Using Deep Belief Network. IEEE Access. 2021;9:135542–9.

39. Yashfi SY, Islam MA, Pritilata, Sakib N, Islam T, Shahbaaz M, et al. Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms. In: 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020. Institute of Electrical and Electronics Engineers Inc.; 2020.

40. Ghosh P, Javed Mehedi Shamrat FM, Shultana S, Afrin S, Anjum AA, Khan AA. Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm. In: Proceedings - 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAI-NLP 2020. Institute of Electrical and Electronics Engineers Inc.; 2020.

41. Manonmani M, Balakrishnan S. An ensemble feature selection method for prediction of CKD. In: 2020 International Conference on Computer Communication and Informatics, ICCCI 2020. Institute of Electrical and Electronics Engineers Inc.; 2020.

42. Vinothini A, Baghavathi Priya S. Design of chronic kidney disease prediction model on imbalanced data using machine learning techniques. Indian Journal of Computer Science and Engineering. 2020 Nov 1;11(6):708–18.

43. Kriplani H, Patel B, Roy S. Prediction of chronic kidney diseases using deep artificial neural network technique. In: Lecture Notes in Computational Vision and Biomechanics. Springer Netherlands; 2019. p. 179–87.

44. Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar. Prediction of Chronic Kidney Disease Using Machine Learning Algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2018 Oct;7(10).

45. Simeone O. A brief introduction to machine learning for engineers. Vol. 12, Foundations and Trends in Signal Processing. 2018.

46. Khdair H, Dasari NM. Exploring Machine Learning Techniques for Coronary Heart Disease Prediction. International Journal of Advanced Computer Science and Applications. 2021;12(5).

47. Aldahiri A, Alrashed B, Hussain W. Trends in Using IoT with Machine Learning in Health Prediction System. Forecasting. 2021;3(1).

48. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019;7.

49. Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. Heart Disease Prediction using Hybrid machine Learning Model. In: Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021. 2021.

50. Ghiasi MM, Zendehboudi S, Mohsenipour AA. Decision tree-based diagnosis of coronary artery disease: CART model. Comput Methods Programs Biomed. 2020;192.

51. Hafeez MA, Rashid M, Tariq H, Abideen ZU, Alotaibi SS, Sinky MH. Performance improvement of decision tree: A robust classifier using tabu search algorithm. Applied Sciences (Switzerland). 2021;11(15).

52. Ayon SI, Islam MM, Hossain MR. Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. IETE J Res. 2022;68(4).

53. Sinaga LM, Sawaluddin, Suwilo S. Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining. In: IOP Conference Series: Materials Science and Engineering. 2020.

54. Krithiga B, Sabari P, Jayasri I, Anjali I. Early detection of coronary heart disease by using naive bayes algorithm. In: Journal of Physics: Conference Series. 2021.

55. Raizada RDS, Lee YS. Smoothness without Smoothing: Why Gaussian Naive Bayes Is Not Naive for Multi-Subject Searchlight Studies. PLoS One. 2013;8(7).

56. Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. Int J Environ Res Public Health. 2021;18(14).

57. Bowlee J. Logistic Regression for Machine Learning. Machine Learning Mastery. 2016;

58. Chao CM, Yu YW, Cheng BW, Kuo YL. Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree. J Med Syst. 2014;38(10).

59. Gavrishchaka V, Senyukova O, Koepke M. Synergy of physics-based reasoning and machine learning in biomedical applications: Towards unlimited deep learning with limited data. Vol. 4, Advances in Physics: X. 2019.

60. Cai Z, Gu J, Wen C, Zhao D, Huang C, Huang H, et al. An Intelligent Parkinson's Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach. Comput Math Methods Med. 2018;2018.

61. Anggoro DA. Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. International Journal of Emerging Trends in Engineering Research. 2020;8(5).

62. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing. 2017;234.

63. Bashar DrA. SURVEY ON EVOLVING DEEP LEARNING NEURAL NETWORK ARCHITECTURES. Journal of Artificial Intelligence and Capsule Networks. 2019;2019(2).

64. Mohapatra S, Swarnkar T, Das J. Deep convolutional neural network in medical image processing. In: Handbook of Deep Learning in Biomedical Engineering: Techniques and Applications. 2020.

65. Sun Y, Xue B, Zhang M, Yen GG. Evolving Deep Convolutional Neural Networks for Image Classification. IEEE Transactions on Evolutionary Computation. 2020;24(2).

66. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev. 2020;53(8).

RESUME

Ahmed Abbas ABD ULSADA graduated first and elementary education in this city. He started undergraduate program in Madenat Alelem University College Department of Computer Engering in 2014. In 2020, he Moved to turkey to purse a M.Sc. degree in computer engineering at Karabük University.