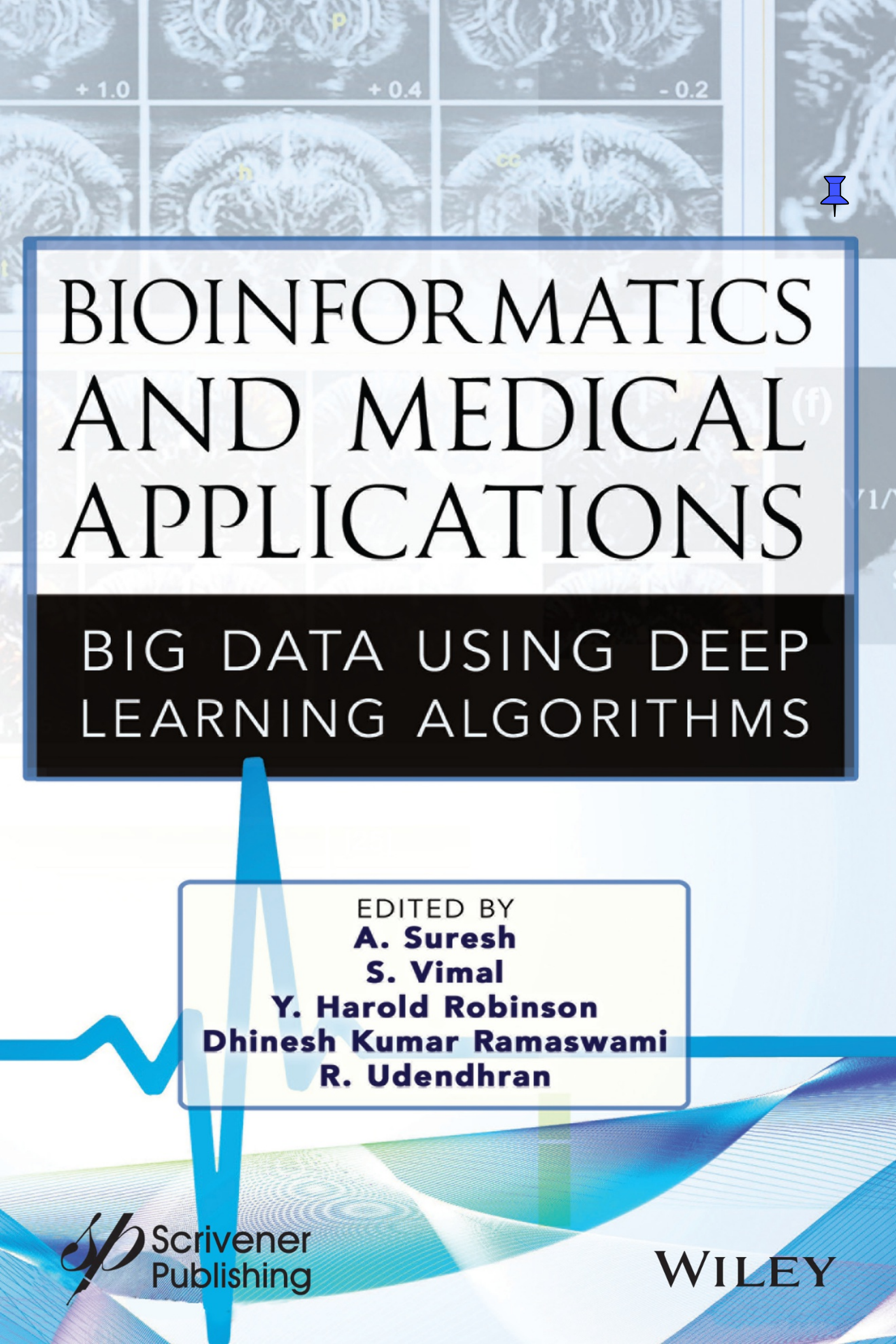

BIOINFORMATICS AND MEDICAL APPLICATIONS

BIG DATA USING DEEP LEARNING ALGORITHMS

EDITED BY
A. Suresh
S. Vimal
Y. Harold Robinson
Dhinesh Kumar Ramaswami
R. Udendhran

Scrivener Publishing

WILEY

Bioinformatics and Medical Applications

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Bioinformatics and Medical Applications

Big Data Using Deep Learning Algorithms

Edited by
A. Suresh
S. Vimal
Y. Harold Robinson
Dhinesh Kumar Ramaswami
and
R. Udendhran

Scrivener
Publishing

WILEY

This edition first published 2022 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-119-79183-6

Cover image: Pixabay.Com
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Contents

Preface

This book features bioinformatics applications in the medical field that employ deep learning algorithms to analyze massive biological datasets using computational approaches and the latest cutting-edge technologies to capture and interpret biological data. In addition to delivering the various bioinformatics computational methods used to identify diseases at an early stage, it also collects cutting-edge resources in a single source designed to enlighten the reader with topics centered on computer science, mathematics, and biology. Since bioinformatics is critical for data management in the current fields of biology and medicine, this book explains the important tools used by bioinformaticians and examines how they are used to evaluate biological data in order to advance disease knowledge.

As shown in the chapter-by-chapter synopsis that follows, the editors of this book have curated a distinguished group of perceptive and concise chapters that reflect the current state of medical treatments and systems and offer emerging solutions for a more personalized approach to the healthcare field. Since applying deep learning techniques for data-driven solutions in health information allows automated analysis, this method can be more advantageous in addressing the problems arising from medical- and health-related information.

- Chapter 1, "Probabilistic Optimization of Machine Learning Algorithms for Heart Disease Prediction," discusses the ensemble learning that overcomes the limitations of a single algorithm, such as bias and variance, by using a multitude of algorithms. It highlights the importance of ensemble techniques in improving the forecast accuracy and displaying an acceptable performance in disease prediction. Additionally, the authors have worked on a procedure to further improve the accuracy of the ensemble method post application by focusing on the wrongly classified records and using probabilistic optimization to select pertinent columns by increasing their weight and doing a reclassification which would result in further improved accuracy.

- Chapter 2, "Cancerous Cells Detection in Lung Organs of the Human Body: IoT-Based Healthcare 4.0 Approach," analyzes three types of cancer—squamous cell carcinoma, adenocarcinoma, and large cell carcinoma—derived from lung tissue, and investigates how AI can customize treatment choices for lung cancer patients.
- Chapter 3, "Computational Predictors of the Predominant Protein Function: SARS-CoV-2 Case," describes the main molecular features of SARS-CoV-2 that cause COVID-19 disease, as well as a high-efficiency computational prediction called the polarity index method. Furthermore, it presents a molecular classification of the RNA-virus and DNA-virus families with results obtained by the proposed non-supervised method focusing on the linear representation of proteins.
- Chapter 4, "Deep Learning in Gait Abnormality Detection: Principles and Illustrations," discusses cerebral palsy, a medical condition which is marked by weakened muscle coordination and other dysfunctions. This chapter proposes a deep learning technique, including support vector machines, multilayer perceptron, vanilla long short-term memory, and bi-directional LSTM, to diagnose cerebral palsy gait.
- Chapter 5, "Broad Applications of Network Embeddings in Computational Biology, Genomics, Medicine, and Health," mainly focuses on the current traditional development of network or graph embedding and its application in computational biology, genomics, and healthcare. As biological networks are very complex and hard to interpret, a significant amount of progress is being made towards a graph or network embedding paradigm that can be used for visualization, representation, interpretation, and their correlation. Finally, to gain more biological insight, further quantification and evaluation of the network embedding technique and the key challenges are addressed.
- Chapter 6, "Heart Disease Classification Using Regional Wall Thickness by Ensemble Classifier," focuses on the cardiac magnetic resonance images that are formed using radio waves and an influential magnetic field to produce images showing detailed structure within and around the heart.

These images can be used to identify cardiac disease through various learning techniques employed to evaluate the heart's anatomy and function in patients. In this chapter, an ensemble classification model is used to classify the type of heart disease.

- Chapter 7, "Deep Learning for Medical Informatics and Public Health," highlights deep learning drawbacks related to data (higher number of features, dissimilar data, reliance on time, unsupervised data, etc.) and model (dependability, understandability, likelihood, scalability) for real-world applications. It emphasizes the DL techniques applied in medical informatics and recent public health case studies related to the application of deep learning and certain critical research questions.
- Chapter 8, "An Insight into Human Pose Estimation and Its Applications," discusses human pose estimation and examines potential deep learning algorithms in great detail, as well as the benchmarking datasets. Recent important deep learning-based models are also investigated.
- Chapter 9, "Brain Tumor Analysis Using Deep Learning: Sensor and IoT-Based Approach for Futuristic Healthcare," proposes an approach for the prediction of brain tumors.
- Chapter 10, "Study of Emission from Medicinal Woods to Curb Threats of Pollution and Diseases: Global Healthcare Paradigm Shift in the 21st Century," focuses on techniques to prevent pollution-related diseases.
- Chapter 11, "An Economical Machine Learning Approach for Anomaly Detection in IoT Environment," presents an improved version of the previous machine learning architecture for ransomware assault in the IoT since it could be more destructive and hence might influence the entire security administration scenario. Therefore, precautions are to be taken to secure the devices as well as data that is being transmitted among themselves, and threats have to be detected at an earlier stage to ensure complete security of the communication. The work proposed in this chapter analyzes the communicating data between these devices and aids in choosing an economically appropriate measure to secure the system.
- Chapter 12, "Indian Science of Yajna and Mantra to Cure Different Diseases: An Analysis Amidst Pandemic with a

Simulated Approach," discusses deep Yagya training, which is an amazingly practical application that is easy to use and exciting, and has a great impact on delicate thinking and emotions.

- Chapter 13, "Collection and Analysis of Big Data from Emerging Technologies in Healthcare," discusses the fact that new diseases, such as COVID-19, are constantly being discovered. Since this results in a tremendous surge in data being generated and a huge burden falling on medical personnel, this is an area in which automation and emerging technologies can contribute significantly. Since combining big data with emerging healthcare technologies is the need of the hour, this chapter focuses on the collection of big data using emerging technologies like radio frequency identification (RFID), wireless sensor networks (WSN), and the internet of things (IoT), and their applications in the medical field. After discussing different data analysis approaches, the challenges and issues that arise during data analysis are explored and current research trends in the field are summarized.
- Chapter 14, "A Complete Overview of Sign Language Recognition and Translation Systems," discusses the use of human body pose and hand pose estimation. Sign language recognition has been conventionally performed by some preliminary sensors and later evolved to various advanced deep learning-based computer vision systems. This chapter deals with the past, present, and future of sign language recognition systems. Sign language translation is also briefly discussed, providing insights into the natural language processing techniques used to accurately convert sign language to translated sentences.

The editors thank the contributors most profoundly for their time and effort.

A. Suresh
S. Vimal
Y. Harold Robinson
Dhinesh Kumar Ramaswami
R. Udendhran
February 2022

1

Probabilistic Optimization of Machine Learning Algorithms for Heart Disease Prediction

Jaspreet Kaur[1]*, Bharti Joshi[2] and Rajashree Shedge[2]

[1]Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India
[2]Department of Computer Engineering Ramrao, Adik Institute of Technology Nerul, Navi Mumbai, India

Abstract

Big Data and Machine Learning have been effectively used in medical management leading to cost reduction in treatment, predicting the outbreak of epidemics, avoiding preventable diseases, and, improving the quality of life.

Prediction begins with the machine learning patterns from several existing known datasets and then applying something very similar to an obscure dataset to check the result. In this chapter, we investigate Ensemble Learning which overcomes the limitations of a single algorithm such as bias and variance by using a multitude of algorithms. The focus is not solely increasing the accuracy of weak classification algorithmic programs however additionally implementing the algorithm on a medical dataset wherever it is effectively used for analysis, prediction, and treatment. The consequence of the investigation indicates that ensemble techniques are powerful in improving the forecast accuracy and displaying an acceptable performance in disease prediction. Additionally, we have worked on a procedure to further improve the accuracy post applying ensemble method by focusing on the wrongly classified records and using probabilistic optimization to select pertinent columns by increasing their weight and doing a reclassification which would result in further improved accuracy. The accuracy hence achieved by our proposed method is, by far, quite competitive.

***Keywords*:** Kaggle dataset, machine learning, probabilistic optimization, decision tree, random forest, Naive Bayes, K means, ensemble method, confusion matrix, probability, Euclidean distance

**Corresponding author*: jaspreetseera@gmail.com

A. Suresh, S. Vimal, Y. Harold Robinson, Dhinesh Kumar Ramaswami and R. Udendhran (eds.)
Bioinformatics and Medical Applications: Big Data Using Deep Learning Algorithms, (1–28)

1.1 Introduction

Healthcare and biomedicine are increasingly using big data technologies for research and development. Mammoth amount of clinical data have been generated and collected at an unparalleled scale and speed. Electronic health records (EHR) store large amounts of patient data. The quality of healthcare can be greatly improved by employing big data applications to identify trends and discover knowledge. Details generated in the hospitals fall in the following categories.

- Clinical data: Doctor's notes, prescription data, medical imaging reports, laboratory, pharmacy, and insurance related data.
- Patient data: EHRs related to patient admission details, diagnosis, and treatment.
- Machine generated/sensor data: Data obtained from monitoring critical symptoms, emergency care data, web-based media posts, news feeds, and medical journal articles.

The pharmaceutical companies, for example, can effectively utilize this data to identify new potential drug candidates and predictive data modeling can substantially decrease the expenses on drug discovery and improve the decision-making process in healthcare. Predictive modeling helps in producing a faster and more targeted research with respect to drugs and medical devices.

AI depends on calculations that can gain from information without depending on rule-based programming while big data is the type of data that can be supplied to analytical systems so that a machine learning model could learn or, in other words, improve the accuracy of its predictions. Machine learning algorithms is classified in three sorts, particularly supervised, unsupervised, and reinforcement learning.

Perhaps, the most famous procedure in information mining is clustering which is the method of identifying similar groups of data. The groups are created in a manner wherein entities in one group are more similar to each other than to those belonging to the other groups. Although it is an unsupervised machine learning technique, such collections can be used as features in supervised AI model.

Coronary illness, the primary reason behind morbidness and fatality globally, was responsible for more deaths annually compared to any other cause [1]. Fortunately, cardiovascular failures are exceptionally preventable and straightforward way of life alterations alongside early treatment incredibly improves the prognosis. It is, nonetheless, hard to recognize high-risk patients because of the presence of different factors that add to the danger of coronary illness like diabetes, hypertension, and elevated cholesterol. This is where information mining and AI have acted the hero by creating screening devices. These devices are helpful on account of their predominance in pattern recognition and classification when contrasted with other conventional statistical methodologies.

For exploring this with the assistance of machine learning algorithms, we gathered a dataset of vascular heart disease from Kaggle [3]. It consists of three categories of input features, namely, objective consisting of real statistics, examination comprising of results of clinical assessment, and subjective handling patient related information.

Based on this information, we applied various machine learning algorithms and analyzed the accuracy achieved by each of the methods. For this report, we have used Naive Bayes, Decision Tree, Random Forest, and various combinations of using these algorithms in order to further improve the accuracy. Numerous scientists have just utilized this dataset for their examination and delivered their individual outcomes. The target of gathering and applying methods on this dataset is to improve the precision of our model. For this reason, we gave different algorithms a shot on this dataset and successfully improved the accuracy of our model.

We suggested using the ensemble method [2] which is the process of solving a particular computer intelligence problem by strategically combining multiple models, such as classifiers or experts. Additionally, we have take the wrongly classified records by all the methods and tried to understand the reason for wrong classification and modify it mathematically in order to give accurate results and improve model performance continuously.

1.1.1 Scope and Motivation

Exploring different classification and integration algorithms to perceive teams in an exceedingly real-world health record data stored electronically having high dimension capacity and find algorithms that detect clusters within reasonable computation time and ability to scale with increasing data size/features while giving the highest possible accuracy. Diagnosis is a challenging process that, as of today, involves many human-to-human interactions. A machine would increase the speed of giving a diagnosis and lead to a more rapid treatment decision and would be able to detect rare events easier than humans.

1.2 Literature Review

Over the years, many strategies have been used regarding data processing and model variability in the field of cardiovascular diagnostics. Authors in [4] show that splitting the data into 70:30 ratio using for tutoring and examination purpose and 10-fold cross proofing putting logistic regression into operation improved the accuracy of the UCI dataset to 87%.

Authors in [5] have used ensemble classification techniques using multiple classifiers followed by score level ensemble for improving the prediction accuracy. They pointed out that maximum voting produces the highest level of development. This functionality is enhanced by using feature selection.

Hybrid approach has been proposed in [6] by consolidating Random Forest along with Linear method leading to a precision of around 90%. In [7], Vertical Hoeffding Decision Tree (VHDT) was used accuracy of 85.43% using 10-fold cross-validation.

Authors in [8] outline a multi-faceted voting system that can anticipate the conceivable presence of coronary illness in humans. It employs four classifiers which are SGD, KNN, Random Forest, and Logistic Regression and joins them in a consolidated way where group formation is performed by a large vote of the species making 90% accuracy.

The strategy utilized in [9] finds these features by way of correlation which can help enhanced prediction results. UCI coronary illness dataset is used to evaluate the result with [6]. Their proposed model accomplished precision of 86.94% which outflanks Hoeffding tree technique which reported accuracy of 85.43%.

Different classifiers, mainly, Decision Tree, NB, MLP, KNN, SCRL, RBF, and SVM have been utilized in [10]. Moreover, integrated methods of bagging, boosting, and stacking have been applied to the database. The results of the examination demonstrate that the SVM strategy utilizing the boosting procedure outflanks the other previously mentioned techniques.

It was exhibited in [11] after various analyses that, if we increase the feature space of RF algorithm while using forecasts and probability of a tuple to belong to a particular class from Naive Bayes model, then we could increase the precision achieved in identifying the categories, by and large.

Studies in [12] suggested that Naive Bayes gives best result when combined with Random Forest. Also, when KNN is combined with RF or RF+NB, the errors remain same suggesting that it is the dominating method.

Authors in [13] compared the precision of various models in classification of coronary disease taking Kaggle dataset of 70,000 records as input. The algorithms used were Random Forest, Naive Bayes, Logistic Regression, and KNN among whom Random Forest was the winner with an accuracy of 73%.

Creators in [14] have fused the results of the AI examination applied on different informational collections focusing on the CAD illness. Common features are compared and extracted from different datasets, and advanced concepts such as fast decision trees and pruned C4.5 tree are administered on it resulting in higher classification accuracy.

Ensemble Optimization is applied in [15] wherein fuzzy logic is used for extraction of features, Genetic Algorithm for reducing them and Neural Network for classifying them. The results have been tested on a sample of size 30 and accuracy achieved is 99.97%

Based on the detailed research discussed above, we analyze by comparing different strategies suggested by different authors in their respective papers. This helps us to quickly understand where we stand presently with respect to these techniques and how they need to mature further.

1.2.1 Comparative Analysis

Please refer to Table 1.1 to get a comparative study of the methods and understand the strengths and weakness of each. This helped us immensely in designing our prototype.

1.2.2 Survey Analysis

Analyzing the literature, we came to know the scope and limitations of prediction techniques. In present days, heart disease rate has significantly increased and the reason behind deaths in the United States. National Heart, Lung, and Blood Institute states that cardiovascular breakdown is a problem in the typical electrical circuit of the heart and siphoning power.

The incorporation of methodologies with respect to information enhancement and model variability has been coordinating preparing and testing of AI model, Cleveland dataset from the UCI file utilized a ton of time since that is a checked dataset and is generally utilized in the preparation and testing of ML models. It has 303 tuples and 14 attributes that depend on the factors that are believed to be associated with an increased risk of cardiovascular illness. Additionally, the Kaggle dataset of coronary illness containing records of 70,000 and 12 patient attributes is also used for the purpose of training and assessment.

Table 1.1 Comparative analysis of prediction techniques.

Title	Problem	Solution	Result
"Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease" [4]	Inspect and look at the precision of four diverse AI calculations which take ROC curve for anticipating and diagnosing cardiovascular ailment analyzing the 14 indicators of the UCI Cardiac Dataset.	Logistic regression, support vector machine, stochastic gradient boosting, and random forest are applied on UCI dataset and accuracy was compared using ROC curve.	Ten-fold cross-validation applied to maximize ROC. Logistic regression performs the best with 87% accuracy.
"Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques" [5]	1. Increase the efficiency of weak classification algorithms. 2. Usage on clinical dataset to show utility to foresee illness at beginning stage.	Research is done on ensemble techniques such as bagging, boosting, majority vote, and stacking, and results are assessed. They are further upgraded by using feature selection.	1. Majority voting produces highest improvement in accuracy. 2. Feature FS2 along with majority voting yields best results.
"Effective heart disease prediction using hybrid machine learning techniques" [6]	Improve precision in forecast of cardiovascular illness	Presented a method called the Hybrid Random forest with Linear Model (HRFLM). It utilizes ANN with back propagation taking as input 13 clinical features	HRFLM ended up being quite precise in the prediction of heart illness.

(*Continued*)

Table 1.1 Comparative analysis of prediction techniques. (*Continued*)

Title	Problem	Solution	Result
"A classification for patients with heart disease based on Hoeffding tree" [7]	Characterize information for patients with coronary sickness and assessment of models used to foresee coronary disease patients.	Hoeffding tree deals with increasing tree proofs and the capacity to gain from steam of huge information assuming that the distribution sample remains constant with time.	Results exhibit an accuracy of around 85% and the processing error value of 14%.
"Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method" [8]	Give more certainty and precision to the Specialist's analysis considering the face that the model is prepared using real information of healthy and sick patients.	Data was divided in 80:20 ratio for training and testing and a combination of four algorithms (SGD, KNN, RF, and LR) was used by majority voting method.	A precision of 90% was achieved based on the hard voting ensemble model.
"Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble Learning Model" [9]	Coronary illness prediction with accessible clinical information is one of the huge difficulties for scientists.	Selected significant attributes by using correlation accompanied with RF and Stratified K-fold cross-validation.	Achieved accuracy of 86.94% which outperforms the 85% precision reported by Hoeffding tree method.

(*Continued*)

Table 1.1 Comparative analysis of prediction techniques. (*Continued*)

Title	**Problem**	**Solution**	**Result**
"A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease" [10]	Compare the accuracy of different data mining classification schemes, employing Ensemble Machine Learning Techniques, for forecasting heart ailments.	Various classifiers, namely, DT, NB, MLP, KNN, SCRL, RBF, and SVM, have been employed.	SVM method using the boosting technique outperforms the other aforementioned methods.
"Increasing Diversity in Random Forests Using Naive Bayes" [11]	Improve the classification accuracy.	Enhanced variety of Random Forests put forward that was constructed by pseudo randomly picking up certain attributes and incorporating Naive Bayes estimation into the training and segregation category.	Proposed method works more efficiently in comparison to other advanced ensemble methods.
"Improved Classification Techniques by Combining KNN and Random Forest with Naive Bayesian Classifier" [12]	Increase classification accuracy.	Utilized average class probabilities to concatenate Naive Bayes, KNN, and Random Forest.	Naive Bayes combined with Random Forest has ended up being the ideal blend.

(*Continued*)

Table 1.1 Comparative analysis of prediction techniques. (*Continued*)

Title	Problem	Solution	Result
"Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data" [13]	Examination of ML models on forecast of cardiovascular illness utilizing patients' cardiovascular hazard factors.	Used Cross Industry Standard Process for Data Mining and four algorithms, namely, RF, NB, LR, and KNN, were used.	Random Forest outperforms other models by achieving an accuracy of 73%, sensitivity of 65%, and specificity of 80%.
"Feature Analysis of Coronary Artery Heart Disease Data Sets" [14]	Combine results of the AI examination applied on various datasets centering on CAD.	Common features are compared and extracted from different datasets and fast decision trees and pruned C4.5 tree are administered on it.	Precision of the collected dataset is around 80%.
"Cardio Vascular Disease Classification Ensemble Optimization Using Genetic Algorithm and Neural Network"[15]	To construct the detection system based on fuzzy logic algorithm for extraction of features making use of neural network classifier of heart disease.	Dataset is categorized via the usage of fuzzy logic, genetic algorithm, and, moreover, training is performed by neural network by the extracting features.	The accuracy is elevated up to 99.97% and the error rate is decreased to 0.987%.

Experimental testing and the use of AI indicate that supervised learning is certain calculation exceeds an alternate calculation for a particular issue or for a specific section of the input dataset; however, it is not phenomenal to discover an independent classifier that accomplishes excellent performance the domain of common problems.

Ensembles of classifiers are therefore produced using many techniques such as the use of separate subset of coaching dataset in a sole coaching algorithm, utilizing distinctive coaching on a solitary coaching algorithm or utilizing multiple coaching strategies. We learnt about the various techniques employed in ensemble method like bagging, boosting, stacking, and majority voting and their affect on the performance improvement.

We also learned about Hoeffding Tree which is the first distributed algorithm for studying decision trees. It incorporates a novel way of dissecting decision trees with vertical parallelism. The development of effective integration methods is an effective research field in AI. Classifier ensembles are by and large more precise than the individual hidden classifiers. This is given the fact that several learning algorithms use local optimization methods that can be traced to local optima.

A few methodologies find those features by relationship which can help successful predictive results. This used in combination with ensemble techniques achieves best results. Various combinations have been tried and tested and none is the standardized/best approach. Each technique tries to achieve a better accuracy than the previous one and the race continues.

1.3 Tools and Techniques

Machine learning and information gathering utilizes ensembles on one or more learning algorithms to get different arrangement of classifiers with the ability to improve performance. Experimental studies have time and again proven that it is unusual to get one classifier which will perform the best on the general problem domain. Hence, ensemble of classifiers is often produced using any of the subsequent methods.

- Splitting the data and using various chunks of the training data for single machine learning algorithm.
- Training one learning algorithm using multiple training parameters.
- Using multiple learning algorithms.

Key ideas such as the data setup, data classification, data mining models, and techniques are described below.