



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

Table of Contents

[Cover](#)

[Title Page](#)

[Copyright](#)

[Preface](#)

[1 Probabilistic Optimization of Machine Learning Algorithms for Heart Disease Prediction](#)

[1.1 Introduction](#)

[1.2 Literature Review](#)

[1.3 Tools and Techniques](#)

[1.4 Proposed Method](#)

[1.5 Conclusion](#)

[References](#)

[2 Cancerous Cells Detection in Lung Organs of Human Body: IoT-Based Healthcare 4.0 Approach](#)

[2.1 Introduction](#)

[2.2 Literature Review](#)

[2.3 Proposed Systems](#)

[2.4 Experimental Results and Analysis](#)

[2.5 Novelties](#)

[2.6 Future Scope, Limitations, and Possible Applications](#)

[2.7 Recommendations and Consideration](#)

[2.8 Conclusions](#)

[References](#)

[3 Computational Predictors of the Predominant Protein Function: SARS-CoV-2 Case](#)

[3.1 Introduction](#)

[3.2 Human Coronavirus Types](#)

[3.3 The SARS-CoV-2 Pandemic Impact](#)

[3.4 Computational Predictors](#)

[3.6 Future Implications](#)

[3.7 Acknowledgments](#)

[References](#)

[4 Deep Learning in Gait Abnormality Detection: Principles and Illustrations](#)

[4.1 Introduction](#)

[4.2 Background](#)

[4.3 Related Works](#)

[4.4 Methods](#)

[4.5 Conclusion and Future Work](#)

[4.6 Acknowledgments](#)

[References](#)

[5 Broad Applications of Network Embeddings in Computational Biology, Genomics, Medicine, and Health](#)

[5.1 Introduction](#)

[5.2 Types of Biological Networks](#)

[5.3 Methodologies in Network Embedding](#)

[5.4 Attributed and Non-Attributed Network Embedding](#)

[5.5 Applications of Network Embedding in Computational Biology](#)

[5.6 Limitations of Network Embedding in Biology](#)

[5.7 Conclusion and Outlook](#)

[References](#)

6 Heart Disease Classification Using Regional Wall Thickness by Ensemble Classifier

6.1 Introduction

6.2 Related Study

6.3 Methodology

6.4 Implementation and Result Analysis

6.5 Conclusion

References

7 Deep Learning for Medical Informatics and Public Health

7.1 Introduction

7.2 Deep Learning Techniques in Medical Informatics and Public Health

7.3 Applications of Deep Learning in Medical Informatics and Public Health

7.4 Open Issues Concerning DL in Medical Informatics and Public Health

7.5 Conclusion

References

8 An Insight Into Human Pose Estimation and Its Applications

8.1 Foundations of Human Pose Estimation

8.2 Challenges to Human Pose Estimation

8.3 Analyzing the Dimensions

8.4 Standard Datasets for Human Pose Estimation

8.5 Deep Learning Revolutionizing Pose Estimation

8.6 Application of Human Pose Estimation in Medical Domains

8.7 Conclusion

References

9 Brain Tumor Analysis Using Deep Learning: Sensor and IoT-Based Approach for Futuristic Healthcare

9.1 Introduction

9.2 Literature Survey

9.3 System Design and Methodology

9.4 Novelty in Our Work

9.5 Future Scope, Possible Applications, and Limitations

9.6 Recommendations and Consideration

9.7 Conclusions

References

10 Study of Emission From Medicinal Woods to Curb Threats of Pollution and Diseases: Global Healthcare Paradigm Shift in 21st Century

10.1 Introduction

10.2 Literature Survey

10.3 The Methodology and Protocols Followed

10.4 Experimental Setup of an Experiment

10.5 Results and Discussions

11 An Economical Machine Learning Approach for Anomaly Detection in IoT Environment

11.1 Introduction

11.2 Literature Survey

11.3 Proposed Work

11.4 Analysis of the Work

11.5 Conclusion

References

12 Indian Science of Yajna and Mantra to Cure Different Diseases: An Analysis Amidst Pandemic With a Simulated Approach

[12.1 Introduction](#)

[12.2 Literature Survey](#)

[12.3 Methodology](#)

[12.4 Results and Discussion](#)

[12.5 Interpretations and Analysis](#)

[12.6 Novelty in Our Work](#)

[12.7 Recommendations](#)

[12.8 Future Scope and Possible Applications](#)

[12.9 Limitations](#)

[12.10 Conclusions](#)

[12.11 Acknowledgments](#)

[References](#)

[13 Collection and Analysis of Big Data From Emerging Technologies in Healthcare](#)

[13.1 Introduction](#)

[13.2 Data Collection](#)

[13.3 Data Analysis](#)

[13.4 Research Trends](#)

[13.5 Conclusion](#)

[References](#)

[14 A Complete Overview of Sign Language Recognition and Translation Systems](#)

[14.1 Introduction](#)

[14.2 Sign Language Recognition](#)

[14.3 Dataset Creation](#)

[14.4 Hardware Employed for Sign Language Recognition](#)

[14.5 Computer Vision-Based Sign Language Recognition and Translation Systems](#)

[14.6 Sign Language Translation System—A Brief Overview](#)

[14.7 Conclusion](#)

[References](#)

[Index](#)

[End User License Agreement](#)

List of Illustrations

Chapter 1

[Figure 1.1 Heatmap of input attributes.](#)

[Figure 1.2 Age distribution.](#)

[Figure 1.3 Presence of cardiovascular disease.](#)

[Figure 1.4 Cholesterol type distribution.](#)

[Figure 1.5 Gender distribution.](#)

[Figure 1.6 Random forest algorithm.](#)

[Figure 1.7 Ensemble methods.](#)

[Figure 1.8 NB confusion matrix.](#)

[Figure 1.9 RF confusion matrix.](#)

[Figure 1.10 DT confusion matrix.](#)

[Figure 1.11 ROC curve analysis.](#)

[Figure 1.12 Proposed architecture.](#)

Chapter 2

[Figure 2.1 Framework of the experimental study of lung cancer stratification.](#)

[Figure 2.2 Sample images of correctly classified and misclassified carcinoma.](#)

[Figure 2.3 More sample images of correctly classified and misclassified carcinom...](#)

Chapter 3

[Figure 3.1 Relative frequency distribution of proteins that express the four SAR...](#)

[Figure 3.2 Zoom over the Figure 3.1. The X-axis represents the five polar intera...](#)

[Figure 3.3 Histograms SARS-CoV-2 structural proteins.](#)

Chapter 4

[Figure 4.1 Cyclic nature of human gait \(Vaughan *et al.* \[9\]\).](#)

[Figure 4.2 LSTM unit \(Bouktif *et al.* \[2\]\).](#)

[Figure 4.3 Bidirectional LSTM \(Yulita *et al.* \[10\]\).](#)

Chapter 5

[Figure 5.1 Network modeling approach that simplifies multi-omics data from the g...](#)

[Figure 5.2 Classification of network embedding and algorithms used \[17-47\].](#)

[Figure 5.3 Various network embedding tools and their applications \[16\].](#)

[Figure 5.4 Taxonomy of biological network alignment.](#)

[Figure 5.5 Illustration of data and methods used in analysis of Adverse Drug Rea...](#)

[Figure 5.6 Illustration of data and methods used in multi-omics data analysis.](#)

Chapter 6

[Figure 6.1 System architecture.](#)

[Figure 6.2 Localization of heart.](#)

[Figure 6.3 Outcome of segmentation.](#)

[Figure 6.4 Feature extraction.](#)

[Figure 6.5 Ensemble classification—Confusion matrix \(using MLP\).](#)

[Figure 6.6 KNN classification—Confusion matrix.](#)

[Figure 6.7 SVM classifier—Confusion matrix.](#)

[Figure 6.8 XG Boost classifier—Confusion matrix.](#)

[Figure 6.9 Logistic regression classifier—Confusion matrix.](#)

[Figure 6.10 MLP classifier—Confusion matrix.](#)

[Figure 6.11 Random forest classifier—Confusion matrix.](#)

[Figure 6.12 Naïve Bayes classifier—Confusion matrix.](#)

Chapter 7

[Figure 7.1 DL architecture \[2\].](#)

[Figure 7.2 Autoencoders \[36\].](#)

[Figure 7.3 Feedforward RNN \[36\].](#)

[Figure 7.4 Long short-term memory_\(LSTM\)_ \[36\].](#)

[Figure 7.5 CNN architecture \[36\].](#)

[Figure 7.6 Framework of CNN \[36\].](#)

[Figure 7.7 Structure of BM \[36\].](#)

[Figure 7.8 DBN architecture \[36\].](#)

Chapter 8

[Figure 8.1 Different poses are illustrated in the figure with the detected key_p...](#)

[Figure 8.2 Image illustrating different body models \[2\].](#)

[Figure 8.3 Challenges to human pose estimation \[7\].](#)

[Figure 8.4 The figure illustrates 2D pose estimation extracted from an image and...](#)

[Figure 8.5 The image illustrates multi-person pose estimation for different numb...](#)

[Figure 8.6 Object annotated images from Pascal VOC Dataset.](#)

[Figure 8.7 These images are from KTH Multi-view Football Dataset \[18\].](#)

[Figure 8.8 MPII Human Pose Dataset with annotated body_joints.](#)

[Figure 8.9 BBC Pose Dataset with overlaid sign language interpreter.](#)

[Figure 8.10 COCO Dataset with the object classes.](#)

[Figure 8.11 Human3.6M Dataset.](#)

[Figure 8.12 Image correspondence with DensePose.](#)

[Figure 8.13 AMASS Dataset.](#)

[Figure 8.14 Left: pose regression; Right: pose refinement.](#)

[Figure 8.15 Overview of the cascaded architecture.](#)

[Figure 8.16 Results from a convolutional pose machine.](#)

[Figure 8.17 Iterative Error Feedback \(IEF\) mechanism for 2D human pose estimatio...](#)

[Figure 8.18 Stacked hourglass module.](#)

[Figure 8.19 Architecture for HRNet.](#)

[Figure 8.20 Approaching 3D human pose estimation.](#)

[Figure 8.21 Left: Results from DensePose-RCNN; Middle: DensePose COCO Dataset an...](#)

[Figure 8.22 DensePose-R CNN model.](#)

[Figure 8.23 Left: pose estimation without blanket; Right: pose estimation with o...](#)

[Figure 8.24 Combined CNN-RNN model.](#)

Chapter 9

[Figure 9.1 Estimated healthcare IoT device installation \[14\].](#)

[Figure 9.2 System design flow chart.](#)

[Figure 9.3 CNN architecture.](#)

[Figure 9.4 Block diagram.](#)

[Figure 9.5 Various classes of tumors.](#)

[Figure 9.6 Codes used to train our CNN model.](#)

[Figure 9.7 Loading page.](#)

[Figure 9.8 Result displayed.](#)

[Figure 9.9 Result displayed.](#)

[Figure 9.10 About section.](#)

[Figure 9.11 Accuracy score.](#)

[Figure 9.12 Comparison chart.](#)

Chapter 10

[Figure 10.1 Standard level of criteria air pollutants and their sources with hea...](#)

[Figure 10.2 World regional capital city ranking, 2018 \(Website: IQAir\).](#)

[Figure 10.3 Diagrammatic representation of component of Hawan Samagri along with...](#)

[Figure 10.4 Society residents chanting Vedic Mantras and performing community Ya...](#)

[Figure 10.5 People in Indian and South Asian continent celebrating Holi Festival...](#)

[Figure 10.6 CCD image sensor for capturing images \(Website: Wikipedia\).](#)

[Figure 10.7 IR proximity sensor for distance measurement \(Website: Flipkart.com\)...](#)

[Figure 10.8 In Yajna, besides burning material objects, chanting and praying are...](#)

[Figure 10.9 The IoT-based sensors capturing the humidity and temperature data fr...](#)

[Figure 10.10 Measurement of different parameters of AQI on November 16, 2019 \(Ya...](#)

[Figure 10.11 Measurement of different parameters of AQI on November 17, 2019 \(Ya...](#)

[Figure 10.12 Measurement of different parameters of AQI on November 18, 2019 \(Ya...](#)

[Figure 10.13 Measurement of different parameters of AQI on November 19, 2019 \(Ya...](#)

[Figure 10.14 Comparative analysis of emission of different gaseous elements and ...](#)

[Figure 10.15 Comparative analysis of averaged environmental parameters in fume e...](#)

[Figure 10.16 Study of the comparative analysis of stability of the environment b...](#)

[Figure 10.17 Collective recital of mantra helps in depression treatment and slee...](#)

Chapter 11

[Figure 11.1 Analysis w.r.t. attribute-transformation clusters.](#)

[Figure 11.2 Analysis w.r.t. rate-transformation clusters.](#)

Chapter 12

[Figure 12.1 Benefits of Yagya or Yagyopathy \[56\].](#)

[Figure 12.2 Graphical analysis for fasting blood sugar with the respect to the a...](#)

[Figure 12.3 Graphical presentation for FBST PRANDAL blood sugar with the respect...](#)

[Figure 12.4 Graphical representation to compare the sugar level results in 4-mon...](#)

[Figure 12.5 Graphical representation to compare the FVC and FEV1 on parameters.](#)

[Figure 12.6 Graphical representation to compare subject's data on different date...](#)

[Figure 12.7 Radiation variation of different electronic gadgets of subjects \(lap...](#)

[Figure 12.8 Left-hand analysis of happiness index of different subjects.](#)

[Figure 12.9 Right-hand analysis of happiness index of different subjects.](#)

[Figure 12.10 Age vs. happiness index \(right hand\).](#)

[Figure 12.11 Age vs. happiness index \(left hand\).](#)

Chapter 13

[Figure 13.1 Taxonomy.](#)

[Figure 13.2 Three-layer architecture for remote health monitoring.](#)

Chapter 14

[Figure 14.1 The Myo armbands usage for Sign Language Recognition.](#)

[Figure 14.2 Smart ring + watch setup to detect movements of the finger during th...](#)

[Figure 14.3 Figure shows the start, middle, and end frames along with the motion...](#)

[Figure 14.4 Images showing annotations of various images extracted from ASLLVD a...](#)

[Figure 14.5 Image showing variation in the training and test set images. The var...](#)

[Figure 14.6 Image showing the capturing system used in the SMILE Swiss German Si...](#)

[Figure 14.7 Image showing the contents of SMILE Swiss German Sign Language Datas...](#)

[Figure 14.8 Figure representing the recording procedure for the SIGNUM corpus \[1...](#)

[Figure 14.9 The figure represents the 32 alphabets and the number of images obta...](#)

[Figure 14.10 This table summarizes some glove sensor systems among which systems...](#)

[Figure 14.11 Image showing the 3D reconstruction through Kinect-based images. Im...](#)

[Figure 14.12 Figure representing usage of multi-feature extraction from an image...](#)

[Figure 14.13 Capsule network for the Sign Language Recognition task \[29\].](#)

[Figure 14.14 Real-Time Sign Language Recognition system using Pose Estimation an...](#)

[Figure 14.15 Network showing the feature extraction and the usage of LSTM to per...](#)

[Figure 14.16 Difference between a Sign Language Recognition and a Sign Language ...](#)

[Figure 14.17 Image showing the Neural Sign Language Translation by employing seq...](#)

[Figure 14.18 Image showing the different human body parts take into account and ...](#)

List of Tables

Chapter 1

[Table 1.1 Comparative analysis of prediction techniques.](#)

[Table 1.2 Dataset attributes.](#)

Chapter 4

[Table 4.1 Confusion matrix for binary classification.](#)

[Table 4.2 Performance analysis of different classifiers.](#)

Chapter 5

[Table 5.1 List of major biological network data resources.](#)

Chapter 6

[Table 6.1 Details of the extracted features.](#)

[Table 6.2 Accuracy score of classification models.](#)

Chapter 8

[Table 8.1 The table contains a list of datasets in 2D and 3D human pose estimati...](#)

Chapter 11

[Table 11.1 Parameters used in simulation.](#)

Chapter 12

[Table 12.1 Yagya decreases air pollution \[60\].](#)

[Table 12.2 Comparison of results.](#)

[Table 12.3 Yagyopathy experiment on diabetic patient.](#)

[Table 12.4 Final result.](#)

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



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

© 2022 Scrivener Publishing LLC

For more information about Scrivener publications please visit

www.scrivenerpublishing.com.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

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

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 Joshi² and Rajashree Shedge²

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

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 morbidity 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