ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

A STUDY FOR DIAGNOSING AND TREATMENT OF CHRONIC AND RECURRENT DISEASES USING DATA MINING TECHNIQUES

Thesis submitted to

SEACOM SKILLS UNIVERSITY

Kendradangal, Bolpur

Dist: Birbhum, PIN - 731 236, West Bengal

In partial fulfillment of the requirements
For the award of the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER APPLICATION

BY

AVIJIT KUMAR CHAUDHURI
Research Scholar, School of Computer Application,
Seacom Skills University
Registration No. – 1711404061915

A STUDY FOR DIAGNOSING AND TREATMENT OF CHRONIC AND RECURRENT DISEASES USING DATA MINING TECHNIQUES

Thesis

Submitted to

SEACOM SKILLS UNIVERSITY

Kendradangal, Bolpur
Dist: Birbhum, PIN - 731 236, West Bengal

In partial fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER APPLICATION

By

AVIJIT KUMAR CHAUDHURI

Research Scholar, School of Computer Application
Registration No. – 1711404061915
SEACOM SKILLS UNIVERSITY
Kendradangal, Bolpur
Dist: Birbhum, PIN - 731 236, West Bengal

SUPERVISED BY

PROF. (DR.) DILIP KUMAR BANERJEE

Professor, Department of Computer Application
SEACOM SKILLS UNIVERSITY
Kendradangal, Bolpur
Dist: Birbhum, PIN - 731 236, West Bengal

AND CO-SUPERVISED BY

PROF. (DR.) ANIRBAN DAS

Professor, Department of Computer Science
UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

A Study for Diagnosing and Treatment of Chronic and Recurrent Diseases using Data Mining Techniques

Abstract

The healthcare industry is one of the most information-intensive sectors. Medical facts, expertise, and information continue to expand every day. It was projected that five terabytes of data per year may be produced in an acute care hospital (Huang et al. 1996). Such data may be used to collect important healthcare information. The tremendous growth in the field of information technology, software development, and system integration technology has been incorporated to produce a new generation of the complex computer system. Information technology researchers are facing challenges to keep pace with these new eras of evolutions. These Challenges are found in the fields of compatibility among heterogeneous systems, securities and privacy issues, management of systems, distribution of data, and reusing and benefiting from the existing resources and data.

The Healthcare system is an example of such (complex computer system) complicated system. In recent times the interest to exercise the growth of communication technology and data mining techniques has increased. So many countries are now changing the approach of regulating individual health protection systems towards a global health protection system across the country by setting up a universal health protection system in communication and building electronic healthcare records.

The Electronic Health Record (EHR) is an organized agglomeration of electronic health data of individual patients or society. It is adequately being shared across healthcare providers in certain states or the country (Gunter & Terry 2005). Health records are a collection of general medical records, patient examinations, patient treatments, medical history, allergies, immunization status, laboratory results, radiology images, and some useful information for examination. Researchers can examine and diagnose various diseases by collecting rich information generated from the analysis of these health records.

The use of EHRs will enable a reduction in expenditure incurred from the use of traditional systems, improving the quality of health care, reducing the allied risk, and accelerate the mobility of records. But the only disadvantage is to resolve the issues of privacy and security in such models for patients which have to be handled by the respective government.

Researchers are inspired by the advancement of EHRs to share the information of electronic healthcare system where the components of the legacy healthcare systems (facilities, workforce, the providers of therapeutics, and education and research institution) appear in a well-organized form and convey the patient's data across the public infrastructure throughout the country.

India is moving rapidly forward in the field of electronic health care information systems across the country. This fast development will produce a huge EHR for the Indian community and healthcare providers, and these patients along with their disease-related information and data can be a valuable resource for future analysis. Therefore, the main aspiration of our present work is to investigate the aspects of utilizing health data for the benefit of humans by using novel machine learning and data mining techniques.

The objective is to design automated methods for the diagnosis of diseases based on the previous data and information collected from the analysis of those data. However, there are definite complications identified with effectively applying this previously acquired patient data, which can make any electronic healthcare system less efficient. These complications are the disputes arising at the time of handling the missing values and how to process them, the issue of innumerable features and attributes to select the most construction components, and the problem of extracting accurate diagnostic markers that can predict the onset of the disease in advance, and the control of different stages of the disease.

In India and all over the world, people are suffering from limited medical resources and long waiting times to receive medical services. According to a Lancet study (NCBI), India ranks 145th out of 195 nations in terms of healthcare quality and accessibility, behind neighbors such as China, Bangladesh, Sri Lanka, and Bhutan. The study showed that India did not work well on the treatment of, for example, tuberculosis, rheumatic heart diseases,

ischemic heart diseases, stroke, testicular cancer, colon cancer, and chronic kidney disease among others (NCBI).

The increasing population of India, the aging population, the modern lifestyle, the climate change, and the new diseases that come into view have presented challenges for the Indian health organizations and state governments to set procedures and plans to manage and cope with the available medical resources, infrastructure, and to deliver decent healthcare services for residents despite the shortages in medical personnel and equipment. In addition, medical services are essential for all individuals and it is the nation's responsibility to develop and sustain the medical infrastructures and services for all residents and citizens. In addition to the shortages in medical personnel and technology, incidents of prescription errors have been increasingly causing minor to major problems for patients. For example, serious health problems may occur because of Adverse Drug Effects (ADE). ADE is caused by mistaken prescription, errors in dosage, miscommunication between physicians and pharmacy, dispensing and administering of drugs, and an inappropriate number of drug intake (NCBI). According to one research (NCBI), ADE may be the sixth biggest cause of mortality in the United States, following heart disease, cancer, stroke, respiratory disorders, and traffic accidents. In Australia, the Australian Department of Health and Aging estimated that around 140,000 hospital admissions every year are due to ADE incidents (Gunter & Terry 2005). Those problems may be avoided by a systematic information transfer between different health care providers (hospitals, medical centers, pharmacies, pathologies, etc.).

Another issue that stands for countries including India is the shortages in medical doctors. Table A shows a comparison between India and other by Health and Physicians Per 1,000 people (International Statistics at NationMaster.com). The Table shows 0.69 physicians per 1000 population. The availability of innovative eHealth technologies such as the one proposed in this research, can help alleviate this shortage.

Table A

	COUNTRY	PHYSICIANS (PER 1,000 PEOPLE)
1	Qatar	7.74 physicians
2	Cuba	6.72 physicians
3	Greece	6.17 physicians
4	Spain	5.08 physicians
5	Austria	4.86 physicians
……	……	……
…….	……	……
95	India	0.69 physicians

This prevailing situation in our country, India, coupled with increasing trend and effectiveness of data mining technique in field of medical treatment and health care serves as a motivation to contribute to the world body of knowledge in solving complex problems such as diseases; which is otherwise is in pre-mature stage. However, with the gamut of health care being very large, the study is focused on select critical diseases.

This research attempts to propose a data mining algorithm that enables:

1. To study the extent of use of data mining techniques in health care.
2. To compare the outcome of such applications of data mining techniques.
3. To identify the gaps in existing research related to treatment of critical diseases.
4. To apply integrated and ensemble or hybrid approach that combines the outcomes of un-supervised and supervised learning techniques.
5. To determine the association rules for prediction of critical diseases such as CVD, breast cancer with recurrence and CKD.

Medical informatics plays a critical role in the utilization of clinical information. In such discoveries, pattern recognition is essential for the diagnosis of new diseases and the investigation of various examples found when the classification of data takes place. Computer-based data recovery may help support quality decision-making and to keep away from human error. Although human basic decision-making is frequently ideal, it is

poor when there are huge amounts of information to be classified. Likewise, effectiveness and exactness of choices will diminish when people are put under pressure and immense work. For instance, a specialist needs to inspect 5 patient records; the person will experience them effortlessly. However, if the quantity of records increments from 5 to 50 with a time constraint, it is almost sure that the accuracy with which the doctor delivers the outcomes won't be as high as the ones acquired when he had just five records to be analyzed.

Structured query languages (SQL) are well-known software tools with limited manipulation flexibility, and SQL is excellent for discovering information as long as the user knows exactly what he or she is looking for. The processor gives the user the correct response for the solution after the user delivers the inquiry. We often see situations when the patient has fever and perspiration signs. SQL cannot provide us diagnoses or decisions that are based on the information given as to whether the patient is experiencing headaches or cold.

This leads to the utilization of data mining in medical informatics, the database that is found in the hospitals, to be specific, the hospital information systems (HIS) containing massive amounts of information which incorporates patients information, data from research facilities which continues developing a seemingly endless amount of time. With the assistance of data mining methods, helpful examples of information can be found inside the data, which will be used for further research and assessment of reports. The other inquiry that emerges is the way to classify or aggregate this massive amount of data. Automatic classification is done dependent on similarities present in the data. The automatic classification procedure is possibly demonstrated productive if the end that is drawn by the automatic classifier is adequate to the clinician or the end-user.

Data mining is a critical stage in the discovery of knowledge. It has attracted the curiosity of many in the information sector in recent years (Frawley, Piatetsky-Shapiro & Matheus 1992). Data cleaning, data integration, data selection, data mining pattern detection, and knowledge presentation are iterative steps in the knowledge discovery process. Data mining, in particular, may perform class description, association, classification, clustering, prediction, and time series analysis. In contrast to traditional data analysis, data mining is discovery-driven.

Finally, certain data mining algorithms employ rules, which are essential for categorization. The rules are derived from patterns found in the training data set and retrieved using various data mining methods.

In this study we look at various data-mining tools, as all data are considered as simple data, to perform automatic classification dependent on the testing dataset and furnish exactness in terms of percentage concerning the number of cases in the testing dataset that were ordered effectively.

Critical diseases may include those diseases that lead to most deaths to mankind, the disease that is yet to find a complete remedy (that is a disease with recurrence post-treatment), and diseases that are not fatal but decrease quality of life and led to perennial expenses, being chronic disease. According to the World Health Organization, heart disease is the leading cause of death that occurs almost equally in men and women (Article-A). By the year 2030, 76% of the deaths in the world will be due to non-communicable diseases (NCDs) (Public Health Agency of Canada 2005).

In a country such as India, breast cancer, especially amongst women, is widely detected. As per a global study, the number of cases of breast cancer India is likely to increase phenomenally to two hundred thousand by 2030. The present level is around 1,15,000. According World Health Organization (WHO), there have been around, on an average taken over last five years, seventy-nine thousand deaths of women from breast cancer (Globocan, W. H. O. 2012). Further studies reveal that this disease is the second most common after cervical cancer. It is also that breast cancer affects women in India a decade before it does to those in the western world. The detection services and medication of early years may explain poor endurance to a great extent. The major issue in this type of disease is its recurrence, causing uncertainty in minds of doctors and patient as to the success of the initial treatment and ways to ensure its non-recurrence or in other words the way to complete cure.

A chronic disease is one lasting 3 months or more, by the definition of the U.S. National Center for Health Statistics. Nowadays Chronic diseases become a considerable factor for global morbidity and mortality. As of now the health problems are being considered only for the developed countries. But at present 4 out of 5 chronic disease deaths now occur in

low- and middle-income countries (Public Health Agency of Canada 2005). In India the estimated number of deaths due to chronic diseases was 3.78 million in 1990 (40.4% of all deaths). This figure will rise to an expected 7.63 million in 2020 (66.7% of all deaths) (Public Health Agency of Canada 2005).

Conventionally, in earlier days, health programs for preventing the chronic diseases have mainly concentrated on hypertension, diabetes mellitus and cardiovascular disease (CVD). However, the escalation in the prevalence of chronic kidney disease (CKD) progressing to end-stage renal disease (ESRD) and the consequent economic accountability of renal replacement therapy (RRT) (Grassmann et al. 2005; CDC 2007) in both advanced as well as progressing countries has highlighted the importance of CKD and its risk factors.

The scope of the study is, thus, bounded to the analysis and prediction of disease that cause most pre-mature deaths, i.e., disease that shows recurrence and most common among women, i.e., breast cancer; and a chronic disease i.e., kidney disease. As outlined above, and the issues like privacy and security has not been included.

In this research, EHRs will be used to produce useful patterns and decision support logic for automatic computer-aided diagnosis by using machine learning techniques. For advancing analysis in this project, the well-known, openly accessible for research purposes free datasets have been used. It is anticipated that the unique algorithms and data mining techniques used in these datasets can be extended to real clinical environments by incorporating them into clinical computer-aided diagnosis and decision support systems. Those datasets are treated as trial datasets before incorporating into the proposed methods into real clinical environments.

In this study both un-supervised and supervised learning methods will be used to arrive at the set of association rules that enables prediction of these selected critical diseases.

The aim of this study is to propose a data mining framework thus reduces errors (type 1 & 2) error in prediction of critical diseases such as breast cancer and CKD.

The role of the research work may be twofold: persistence in its importance in clinical trials and research, and perhaps in helping to reclassify young patients at low or moderate risk. It has a key role as a research tool in vascular biology and has helped to further our

understanding of the atherosclerotic disease. By this research work, we will be able to identify dimensions and factors affecting human health. To identify the attributes, causes, and present approaches to particular recurrent and chronic diseases. To carry out a gap analysis on existing approaches to apply Data Mining tools and techniques for developing a diagnostic tool for treating medical diseases.

References

Huang, H., Tsai, W. T., Bhattacharya, S., Chen, X. P., Wang, Y., & Sun, J. (1996, August). Business rule extraction from legacy code. In Proceedings of 20th International Computer Software and Applications Conference: COMPSAC'96 (pp. 162-167). IEEE.

Gunter, T. D., & Terry, N. P. (2005). The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. Journal of medical Internet research, 7(1), e3.

NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4585088/.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57-57.

Article-A. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

World Health Organization, Public Health Agency of Canada, & Canada. Public Health Agency of Canada. (2005). Preventing chronic diseases: a vital investment. World Health Organization.

Globocan, W. H. O. (2012). Estimated cancer incidence, mortality and prevalence worldwide in 2012. Int Agency Res Cancer.

Grassmann, A., Gioberge, S., Moeller, S., & Brown, G. (2005). ESRD patients in 2004: global overview of patient numbers, treatment modalities and associated trends. Nephrology dialysis transplantation, 20(12), 2587-2593.

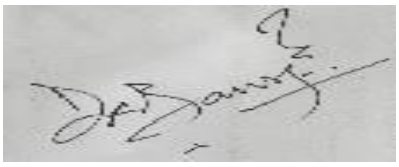
Centers for Disease Control and Prevention (CDC). (2007). Prevalence of chronic kidney disease and associated risk factors—United States, 1999–2004. MMWR Morb Mortal Wkly Rep, 56(8), 161-5.

Supervisor's Certificate

This is to certify that the thesis entitled **A Study For Diagnosing And Treatment Of Chronic And Recurrent Diseases Using Data Mining Techniques** submitted to **Seacom Skills University, Kendradangal, Bolpur, Dist-Birbhum, PIN - 731 236, West Bengal** in partial fulfillment of the requirement for the award of the degree of **Doctor Of Philosophy In Computer Application,** is an authentic and original work carried out by **Mr. Avijit Kumar Chaudhuri** with **Registration No. - 1711404061915** under my supervision.

The matter embodied in this thesis report is genuine work done by the student and has not been submitted whether to this University or to any other University / Institute for the fulfillment of the requirements of any course of study.

SUPERVISOR



Prof. (Dr.) Dilip Kumar Banerjee

Professor Department of Computer Application

Seacom Skills University

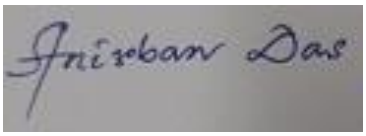
Kendradangal, Bolpur, Dist: Birbhum, PIN - 731 236, West Bengal

Co-Supervisor's Certificate

This is to certify that the thesis entitled **A Study For Diagnosing And Treatment Of Chronic And Recurrent Diseases Using Data Mining Techniques** submitted to **Seacom Skills University, Kendradangal, Bolpur, Dist-Birbhum, PIN - 731 236, West Bengal** in partial fulfillment of the requirement for the award of the degree of **Doctor Of Philosophy In Computer Application,** is an authentic and original work carried out by **Mr. Avijit Kumar Chaudhuri** with **Registration No. - 1711404061915** under my co-supervision.

The matter embodied in this thesis report is genuine work done by the student and has not been submitted whether to this University or to any other University / Institute for the fulfillment of the requirements of any course of study.

CO-SUPERVISOR



Prof. (Dr.) Anirban Das
Professor - Dept of Computer Science
University of Engineering and Management, Kolkata

Acknowledgement

I would like to thank my supervisors, Prof. (Dr.) Dilip K. Banerjee and Prof. (Dr.) Anirban Das, for the patient guidance, encouragement and advice they have provided throughout my time as their student. I have been extremely lucky to have supervisors who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to thank Prof. (Dr.) Suparna Sanyal Mukherjee, for providing guidance and feedback throughout this research work.

Last but not the least I would like to thank my family members and friends for their constant source of inspiration.

AVIJIT KUMAR CHAUDHURI
Registration No. - 1711404061915
Year of Admission: 2017
Dept. – COMPUTER APPLICATION

Table of Contents

List of Publications

Research Paper Published/Accepted in Journals

(1) (2021). Early prediction of heart disease using the most significant features of diabetes by machine learning techniques. Asian Journal For Convergence In Technology (AJCT), 7(1), 168-178.

(2) (2021). Application of data mining techniques for avoiding underestimation of an event. Asian Journal For Convergence In Technology (AJCT), 7(1), 179-189.

(3) (2021). A novel enhanced decision tree model for detecting chronic kidney disease. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 1-22.

(4) (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 100011.

(5) (2020). Early Detection of Cardiovascular Disease in Patients with Chronic Kidney Disease using Data Mining Techniques. Asian Journal For Convergence In Technology (AJCT), 6(3), 65-76.

(6) (2020). An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research. International Journal of Recent Technology and Engineering (IJRTE), 8(6), 1096-1106.

(7) (2020). Identifying the Association Rule to Determine the Possibilities of Cardio Vascular Diseases (CVD). In International Conference on Advanced Machine Learning Technologies and Applications (pp. 219-229). Springer, Singapore.

(8) (2019). Identification of the recurrence of breast cancer by discriminant analysis. In Emerging technologies in data mining and information security (pp. 519-532). Springer, Singapore.

Research Paper Presented in Conferences

(1) (2020). Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases. 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298372.

(2) (2020). Role of Data Mining techniques and MCDM model in detection and severity monitoring to serve as precautionary methodologies against 'Dengue'. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.

List of Tables

List of Figures

Chapter 1

Introduction

Introduction

The Healthcare industry is one of the most information-intensive sectors. Daily, medical information, expertise, and data continue to grow. An acute care hospital is reported to generate five terabytes of data per year (Huang et al. 1996). The crucial point is that these data can be used to extract useful information for quality healthcare. Enormous progress has been implemented in information technology, software development, and system integration engineering to produce a new generation of complex information-intensive systems. Scholars in information technology face challenges in keeping pace with these new stages of evolution.

Such problems are seen in heterogeneous application integration, security and privacy issues, system management, information sharing, and reuse and gain from existing resources and software. The healthcare system is an example of such a complicated system. Recently, there has been an increased interest in the growth of communication technology and data mining techniques. By setting up a universal health protection system in communication and building electronic healthcare records, so many countries are now shifting the approach of regulating individual health protection systems towards a global health protection system across the country. The Electronic Health Record (EHR) is an organized agglomeration of individual patients or society's electronic health data that healthcare providers can share (Gunter & Terry 2005).

Health records are a set of general medical reports, clinical examinations, hospital procedures, history of medications, allergies, immunization status, laboratory results, radiology photos, and some useful information for review. Scientists may analyze and identify different diseases by obtaining rich data from the study of these health records. Uses of EHRs will benefit from reducing traditional system expenditure, improving health care quality, reducing the associated risk, and accelerating record mobility. However, the only downside is to address privacy and security problems in such models for patients that policymakers need to handle.

In health informatics, the use of clinical records plays a vital role. To detect new diseases and study different patterns as data is categorized, identifying patterns in such findings is essential. Computer-supported information recovery can help technical decision-making and human error avoidance. Although the human decision-making process is always good, it is poor where a large volume of data is classified. The consistency and accuracy of decisions will also decline when people are under pressure and in unprecedented working conditions. The job is simpler when a doctor reviews fewer records, say, five case records. However, if the number of records increases from 5 to 50 with a time limit, it is almost certain that the precision with which the doctor produces the findings will not be as good as those obtained when he had to examine just five records.

Researchers were motivated by the development of EHRs to share information about the digital healthcare system in which the components of the traditional healthcare systems (facilities, employees, clinical providers, and educational and research institutions) appear in a well-organized manner and distribute patient data across the public infrastructure across the world. This rapid development will produce an enormous EHR for the Indian community and health care providers, and all this information and data related to patients and diseases can be a valuable resource for future analysis. Therefore, the primary aspect of our present work is to examine the complexities of human use of health data employing modern machine learning methods and data extraction. Its purpose is to design automated methods for diagnosing diseases based on the data obtained in the study previously available. However, there are definite problems that can make any electronic healthcare system less successful by efficiently using these previously obtained patient data. Such concerns include contradictions in the control of and treatment of missed values, different attributes and characteristics, selection of most appropriate components, and acquiring reliable diagnostic markers to estimate the disease's onset and management of multiple phases of the disease.

Structured query languages (SQL) with very little deceptive freedom are quite well-known computing tools, and SQL helps locate information as long as the user knows precise needs. Once the user provides the question, the processor can provide the user with the exact response needed for the solution. We often encounter cases where fever and sweating signs are present in the patient. Based on the details given, SQL does not

give us a diagnosis or a judgment about whether the patient has a headache or a cold. This applies to the use of data mining in medical computer science, the healthcare archive, including the hospital information systems (HIS) containing vast volumes of information that contains patient information, laboratory data that continues to expand year after year. Valuable knowledge patterns can be contained using data mining methods. The other question that emerges is how this large volume of data should be categorized or put together. Automatic classification is conducted based on similarities in the results.

The automatic classification technique is successful only if the clinician or end-user accepts the automatic classifier's interpretation. In this dissertation, we are concerned with text data. Some of these issues may be overcome utilizing a context-based text analysis, such as automatic classification or diagnosis. There are typical methods for extracting features from the data submitted. These features are supported by machine learning using pattern extraction techniques. Typically, these features contain specific patterns or terms that can be used to extract other words or patterns important to the end-user to help classify the results.

However, in our research work, we look at different data mining types and conduct an automated classification based on the test data set; provide consistency within percentage terms for the number of cases in the test data set correctly classified. The basic idea behind this study is to propose a computerized method for the diagnosis of recurrence of breast cancer based on previous evidence and studies as an example, based on the potential of data processing techniques and advanced technology. The study's focus is not limited to the problems listed above, and topics such as privacy and protection have not been discussed. In this research, EHRs will be used as Collaborative Adaptive Health Management Data Sources and will use machine learning techniques to develop usage patterns and decision support logic for automated computer-aided diagnostics. The well-known, freely available data sets for research purposes have been used to advance the study of this research work. The innovative algorithms and data mining techniques used in these databases are meant to be extended to actual clinical environments by incorporating them into computer-aided clinical diagnosis and decision support systems. After the proposed approaches have been implemented into actual clinical environments, these datasets are treated as reference datasets.

In all the case studies discussed in this thesis, we know the types of results of the various cases to concentrate on supervised data mining methods. If no data is available on the classification or outcomes of the incidents, the consequence would be unsupervised learning techniques. While none of the data makes much sense to compiler or machine learning algorithms, the classification and categorization of text data are much easier than other data types. Tests are often more accurate for text information and are obtained more easily than for other forms of data. Applications may be installed on the desktop or portable devices such as PDAs or smartphones, with mobile computing leading the industry. These machines are more useful than notebooks or laptops, which often need easy access. The downside of today's PDAs is its low processing power and limited storage capacity. Because of these factors, it is not feasible to run these algorithms on PDA. Finally, some of the algorithms for data mining use rules that are necessary for categorization. Rules are derived based on trends in the training data set extracted from different data mining algorithms. It is possible to perform this rule-based stage on a computer. These can be stored on a PDA once these rules are collected. Patient inputs can be fed to the PDA, and input identification will occur in real-time based on the rules stored in the system.

In conclusion, the purpose of this research is to use the history, clinical information, and patient databases of the EHRs (Electronic Health Records) to identify indicators for early diagnosis and treatment of cardiovascular disease and breast cancer using an enhanced, intelligent approach consisting of missed processing features. The study is expected to build certain classifiers, templates, and resources that can help physicians detect diseases. To obtain a more reliable diagnosis, the goal is to create an expert method that combines human experience and engineering expertise. This model can assist physicians in decision-making and double-check the medical evaluation (Evidence-based Diagnosis) of disease surveillance by grouping patients into similar health trends for better and more efficient treatment plans. To answer the research goals alluded to above, the research issue is formulated concerning the following questions.

Question 1:

Will the hybridization and stacking of classifiers of current machine learning systems contribute to improved methods for the medical diagnosis of cardiovascular disease and

breast cancer disease in terms of recognition accuracy, noise resistance, and missing values?

Question 2:

How discriminatory dataset features may boost predictions in the case of cardiovascular disease and breast cancer disease, for example.

The key goal is to build models of prediction and research that will enhance healthcare facilities. This work makes developments in the implementation of more effective machine learning techniques in the field of disease identification and prediction. These developments shall include:

- Establishment of innovative models of machine learning and design for the healthcare sector
- Create new approaches to organize and convert training data so that present machine learning algorithms can be applied to real health issues.
- Development and expansion of conventional machine learning methods to the fields of disease identification and analysis, classification or regression enhancement metrics, and/or algorithm performance increase (e.g., training period, prediction period, memory requirements, etc.)

This research does not discuss the abstract complexity of these algorithms in terms of technical definition but does include comparisons of accuracies of various DM techniques for revealing potential discrepancies between optimization of ML algorithms and deployment in healthcare. Only three applications in the area of disease management, fatal diseases (cardiovascular disease), Chronic diseases (chronic kidney disease and diabetes) and recurrent diseases (breast cancer and cervical cancer) are listed, irrespective of the multitude of smart healthcare innovations.

Research Motivation

There are insufficient healthcare services and longer waiting periods in India and around the world for healthcare. According to a Lancet study, India ranks 145th among 195 countries in healthcare quality and accessibility, behind its neighbors such as China, Bangladesh, Sri Lanka, and Bhutan. According to the study, India performed poorly in

tackling cases of tuberculosis, rheumatic heart diseases, Ischemic heart diseases, stroke, testicular cancer, colon cancer, and chronic kidney disease among others.

India's growing population, aging population, the urban lifestyle, climate change, and emerging disease have provided Indian medical institutions and governments with challenges for developing procedures and proposals for maintaining and dealing with the limited medical tools, facilities, and decent healthcare services for people despite the scarcity of resources. Furthermore, medical facilities are important to all people and the nation must build and maintain medical infrastructures and services for all residents. Increases in medication error, in addition to medical professionals and technical shortfalls, create major complications for patients. For example, because of adverse drug effects (ADE), severe health problems can occur. ADE is caused by the incorrect diagnosis, mistakes in dose, physician-pharmacy miscommunication, medication delivery and management, and excessive drug intake[1]. For example, the ADE study[2] found that after coronary disorder, cancer, stroke, lung disease, and road accidents in the USA, ADE is the sixth leading cause of death. In Australia, it is estimated that around 140,000 hospital admissions are a result of ADE events annually in the Australian Ministry of Health & Aging[3]. This can be prevented by the coordinated sharing of knowledge between numerous health care providers (hospitals, medical centers, pharmacies, pathologies, etc.).

The lack of healthcare practitioners is another problem for countries like India. Table 1 shows a comparison of healthcare and doctors per 1000 people among India and others (International Statistics at NationMaster.com). The table shows 0.69 doctors for every 1000 residents compared to more than 2 physician per patient in around 66 countries and one per patient in another 22 countries (out of 150). This scarcity can be relieved by the availability of new technology for eHealth, such as those suggested by this report.

Table 1.1. Comparison of Healthcare and Doctors per 1000 People Among India and Other Countries

	COUNTRY	PHYSICIANS (PER 1,000 PEOPLE)		COUNTRY	PHYSICIANS (PER 1,000 PEOPLE)
1	Qatar	7.74 physicians	76	Colombia	1.47 physicians
2	Cuba	6.72 physicians	77	Panama	1.46 physicians
3	Greece	6.17 physicians	78	China	1.46 physicians
4	Spain	5.08 physicians	79	Syrian Arab Republic	1.46 physicians
5	Austria	4.86 physicians	80	Maldives	1.42 physicians
6	Georgia	4.76 physicians	81	Brunei Darussalam	1.36 physicians
7	Russian Federation	4.31 physicians	82	Tunisia	1.22 physicians
8	Norway	4.16 physicians	83	Algeria	1.21 physicians
9	Switzerland	4.08 physicians	84	Malaysia	1.20 physicians
10	Portugal	3.87 physicians	85	Seychelles	1.19 physicians
11	Kazakhstan	3.86 physicians	86	Trinidad and Tobago	1.18 physicians
12	Australia	3.85 physicians	87	Albania	1.13 physicians
13	Azerbaijan	3.79 physicians	88	Viet Nam	1.11 physicians
14	Belgium	3.78 physicians	89	Chile	1.03 physicians
15	Sweden	3.77 physicians	90	Peru	0.92 physicians
16	Bulgaria	3.76 physicians	91	Iran	0.89 physicians
17	Uruguay	3.74 physicians	92	Belize	0.83 physicians
18	Iceland	3.73 physicians	93	Pakistan	0.83 physicians
19	Czech Republic	3.71 physicians	94	Nicaragua	0.70 physicians
20	Germany	3.69 physicians	95	India	0.69 physicians
21	Israel	3.65 physicians	96	Sri Lanka	0.68 physicians
22	Lithuania	3.64 physicians	97	Morocco	0.62 physicians
23	Latvia	3.54 physicians	98	Iraq	0.61 physicians
24	Belarus	3.51 physicians	99	Tonga	0.56 physicians

25	Italy	3.49 physicians	100	Myanmar	0.50 physicians
26	Denmark	3.48 physicians	101	Samoa	0.48 physicians
27	France	3.45 physicians	102	Bolivia	0.44 physicians
28	Hungary	3.41 physicians	103	Fiji	0.43 physicians
29	Slovakia	3.35 physicians	104	Nigeria	0.40 physicians
30	Estonia	3.34 physicians	105	Thailand	0.39 physicians
31	Ukraine	3.25 physicians	106	Kiribati	0.38 physicians
32	Argentina	3.21 physicians	107	Namibia	0.37 physicians
33	Ireland	3.17 physicians	108	Botswana	0.34 physicians
34	Malta	3.11 physicians	109	Bangladesh	0.29 physicians
35	Republic of Moldova	3.07 physicians	110	Cabo Verde	0.29 physicians
36	Lebanon	3.07 physicians	111	Sudan	0.28 physicians
37	Finland	2.90 physicians	112	Laos	0.27 physicians
38	Netherlands	2.86 physicians	113	Djibouti	0.23 physicians
39	Egypt	2.83 physicians	114	Solomon Islands	0.22 physicians
40	Cyprus	2.77 physicians	115	Guyana	0.21 physicians
41	Luxembourg	2.77 physicians	116	Cambodia	0.21 physicians
42	United Kingdom	2.74 physicians	117	Yemen	0.20 physicians
43	Armenia	2.74 physicians	118	Afghanistan	0.19 physicians
44	New Zealand	2.73 physicians	119	Kenya	0.18 physicians
45	Croatia	2.71 physicians	120	Micronesia (Fed. States of)	0.18 physicians
46	Mongolia	2.67 physicians	121	Madagascar	0.16 physicians
47	TFYR Macedonia	2.63 physicians	122	Côte d’Ivoire	0.14 physicians
48	Jordan	2.56 physicians	123	Indonesia	0.14 physicians
49	Slovenia	2.54 physicians	124	Mauritania	0.13 physicians
50	Uzbekistan	2.54 physicians	125	Uganda	0.12 physicians
51	United Arab Emirates	2.53 physicians	126	Vanuatu	0.12 physicians
52	Kuwait	2.50 physicians	127	Guinea	0.10 physicians

53	United States of America	2.41 physicians	128	Ghana	0.10 physicians
54	Kyrgyzstan	2.40 physicians	129	Congo	0.10 physicians
55	Turkmenistan	2.39 physicians	130	Mali	0.08 physicians
56	Saudi Arabia	2.39 physicians	131	Timor-Leste	0.07 physicians
57	Romania	2.38 physicians	132	Zimbabwe	0.07 physicians
58	Japan	2.30 physicians	133	Zambia	0.07 physicians
59	Serbia	2.11 physicians	134	Benin	0.06 physicians
60	Tajikistan	2.10 physicians	135	Senegal	0.06 physicians
61	Poland	2.07 physicians	136	Papua New Guinea	0.06 physicians
62	Canada	2.07 physicians	137	Rwanda	0.06 physicians
63	Oman	2.05 physicians	138	Togo	0.05 physicians
64	Montenegro	2.03 physicians	139	Burkina Faso	0.05 physicians
65	Republic of Korea	2.02 physicians	140	Guinea-Bissau	0.04 physicians
66	Mexico	1.99 physicians	141	Mozambique	0.04 physicians
67	Singapore	1.92 physicians	142	Gambia	0.04 physicians
68	Libya	1.90 physicians	143	Somalia	0.04 physicians
69	Barbados	1.81 physicians	144	Bhutan	0.02 physicians
70	Brazil	1.79 physicians	145	Sierra Leone	0.02 physicians
71	Bosnia and Herzegovina	1.69 physicians	146	Ethiopia	0.02 physicians
72	Ecuador	1.69 physicians	147	Niger	0.02 physicians
73	El Salvador	1.60 physicians	148	Malawi	0.02 physicians
74	Turkey	1.58 physicians	149	Liberia	0.01 physicians
75	Bahrain	1.49 physicians	150	Tanzania	0.01 physicians

Two findings from The Lancet analysis released by the ESC Congress for 2019, Prospective Urban and Rural Epidemiology (PURE), include new information[4] about the prevalence, hospitalization, and mortality of common illness and[5] modifiable cardiovascular risk factors in adults of middle age in 21 high-income, middle-income and low-income countries (HIC, MIC, LIC) (HIC, MIC, LIC).

A new study released by the Lancet and published in the 2019 ESC Congress[6] indicates that cardiovascular disease (CDV) is still the leading cause of death among middle-aged people and is responsible for 40% of all deaths. However, cancer now has twice as many deaths as CVD[4].

The PURE study is the only large prospective international cohort study that includes substantial data from a large number of MICs and LICs as well as HIC using standardized and concurrent methods of sampling, measuring, and monitoring. The first study, which monitored 162,534 middle-aged adults (age 35-70, 58% female) in 4 HICs, 12 MICs, and 5 LICs over a span of 9.5 years (between 2005-2016), found that CVD-related deaths were 2.5 times more common in middle-aged adults in LICs than in HICs due to the significantly lower burden of CVD-related risk factors compared to wealthier countries.

Authors suggested that higher CVD-related mortality in LIC could be primarily attributable to poorer quality of health care as the first hospitalization and use of CVD treatment were found to be significantly lower in LIC and MIC compared with HIC.

Cancer Incidence

The projected incidence of cancer patients was estimated as 9.6 million worldwide in 2018 and India's share of it was an alarming 8.17%, data recently released by the International Agency for Research on Cancer (WHO) of the World Health Organization reports. Another report released by Lancet within a couple of hours repeated the grim forecast of the heart disease claiming it is the second-largest killer in India. Although statistics from the WHO demonstrate India's position on the global burden for cancer, between 1990 and 2016[7]. The Lancet study compares the prevalence of non-communicable diseases in India.

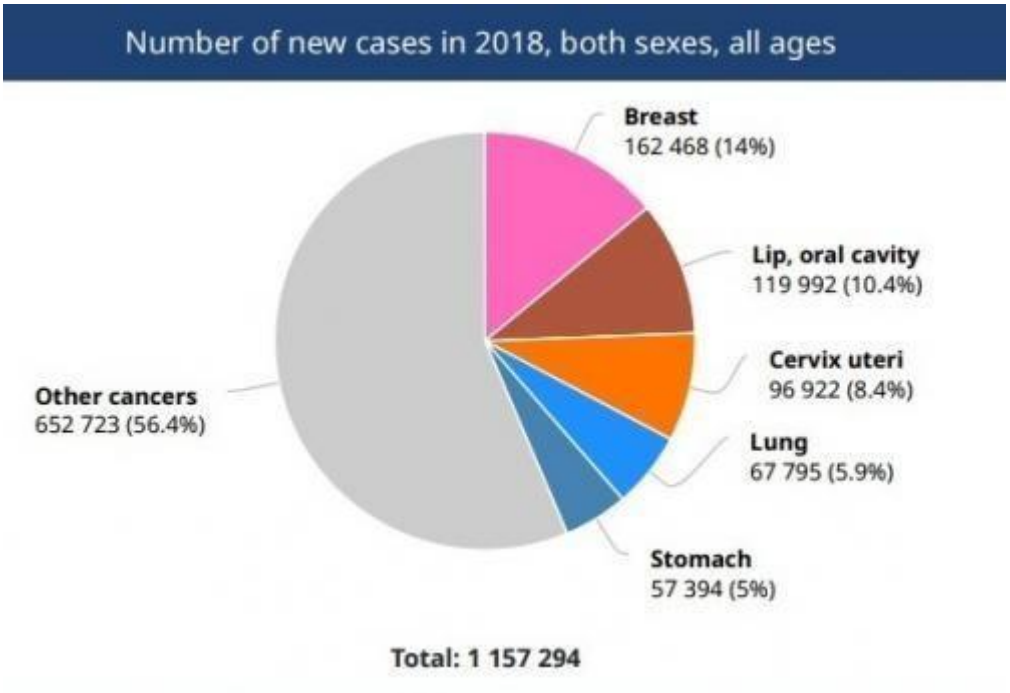


Figure 1.1. IARC Report Showing a Breakup of New Cancer Cases that will be Diagnosed in India in 2018 (Credit: IARC)

The World Health Organization (WHO) reported that one in ten Indians would grow cancer in their lives, and one in fifteen will die of this disease, and an estimated 1.16 million new cancer cases occurred in India in 2018. Therefore, the disease will be prone to more women than men, exactly 17,204. However, between 1990 and 2016, the incidence of breast cancer has increased by 39.1 percent and is the most common cancer among women in India, reflecting the highest raw incidence rate and prevalence of any type of cancer, according to the Lancet report.

IARC also notes that while lip cancer, oral cavity (16.1%), and lung cancer (8.5%) are the most prevalent among Indian men, they are the most prevalent among breast women (27.7%) and uterine cervix (16.5%)[7].

The Lancet report adds that the number of cancer deaths in India increased by 112 percent between 1990 and 2016 and the cancer rate also increased by 48.7% simultaneously. The study also points out that in 2016 the country had 67,000 patients with lung cancer, 72.2 percent of whom were men, as well as a rise in liver cancer of 32.2 percent since 1990, with 30,000 cases reported in 2016[7].

Discovering the disease in its early stages can minimize breast cancer or CVD disease catastrophe. Computing software and machine learning techniques can be used to support doctors in diagnosing and forecasting the illness so they can provide the appropriate measure to prevent the impact, including the risk of death. Breast cancer is 100 times higher for women than men, but for men, it appears to be worse due to diagnostic delays[8]. Breast cancer survival rates significantly differ based on the type of cancer, phase, care, and patient's geographical position. In the Western world, for example, mortality rates are high. But the survival rates are much lower in developing countries.

Moreover, the current trend cannot also be continued indefinitely in the costs of health services. Globally, the care services must improve treatment and reduce costs for patients. Healthcare services have collected a vast volume of organized and unstructured data, but the potential to improve patient care is just starting to be capitalized on. Machine learning delivers better outcomes in the area of healthcare. Machine learning and big data in pharmaceutical and medical applications could generate revenue of up to USD100B per annum, according to the McKinsey report.

Machine learning (ML) in healthcare is used in numerous applications. The application of ML can be typically classified into [9]:

1) Identifying/diagnosing Diseases

2) Customized treatment / behavioral improvement

3) Discovery of Drugs / Production

4) Study on the clinical trial

5) Radiology and Radiotherapy

6) Intelligent patient reports online

7) Prediction of the Epidemic Outbreak

Data mining is a vital phase in the discovery of knowledge. It has been of interest in the information industry in recent years (Fayyad et al. 1996). The method of discovery of knowledge involves an iterative set of data cleaning, integration of data, data selection,

data mining patterns, and the interpretation of knowledge. Data mining can in particular carry out class definitions, associations, classifications, clustering, predictions, and time-series analyses. Data mining is discovery-driven, in contrast to standard data processing.

This study aims to improve diagnosis of diseases for men and women around the world in the fields of breast cancer and cardiovascular disease.

Research Objectives

Literature reveals that the use of machine learning in healthcare fields is popular as it is widely used in today's healthcare sectors. Nevertheless, despite their growing use and their tremendous promise, many areas in healthcare still do not make full use of the latest technologies developed in the field of machine learning. In designing expert systems that can help doctors identify and anticipate diseases in early stages, computational and machine learning techniques can greatly help to solve health care problems. This system can reduce costs, waiting times, and free human experts for more research, as well as eliminate mistakes and errors of medical professionals (Riley & Giarratano 2005). One of the most important medical diagnostic research areas has become a computer-aided diagnosis (CAD) and medical expert systems and tools. CAD's goal is to develop an expert system that blends human experience with engineering intelligence to efficiently achieve a more precise diagnosis. CAD can be used to help physicians identify and anticipate diseases. Physicians can therefore immediately provide the necessary treatment to prevent the loss, including the possibility of death.

It is relatively new to implement artificial intelligence in healthcare. This study aims to show that machine learning can be applied to clinical databases, which identifies the crucial factors causing the diseases and classifies the information with reasonable accuracy. The learning algorithms are provided with a training set, taken from leading databases and Electronic Health Records (EHRs), namely, UCI and Framingham, for a good prediction or classification from which rules or patterns are derived to help identify the testing data set.

In this work, extant widely used tools and data mining algorithms have been to compare the inconveniences and shortcomings.

This study aims to extract the association rules causing the diseases and plausible course of action (for treatment of ailment) thereof.

This research attempts to propose a data mining algorithm that enables:

i. Identification of significant features: This will enable reduction in medical tests, and ease in diagnosis.
ii. Identification of right data mining techniques, as results vary with database-size, dichotomy and number of variables: This will reduce the problem of over and under fitting of data and ensure higher accuracy – precision, specificity, and sensitivity.
iii. Identification of association rules: This will help customized treatment and also aid medical policy makers to decide on medical insurances and health benefits.

The above objectives are aimed at addressing the requirements – diagnosis and treatment – of two diseases, namely, cardiovascular disease (CVD) and breast cancer.

The research finally aims at suggesting a framework for development of an expert-system of healthcare enabling prediction, medical-test recommendations, diagnosis and treatment.

Research Methodology

Introduction

In Western scientific tradition, the Positivist (so-called scientist) and the interpretive (so-called anti-positivist) two main science ideal models were separated (Galliers 1992). However, Dash (Dash 2005) articulated three key forms of science standards: positivism, anti-positivism, and critical thinking. Positivism Worldview depends on perception and thinking as an instrument for understanding a specific issue or behavior.

This worldview typically involves the influence of variables and predictions based on previous perceptions or history. Positivist experts are curious about the consequences of a single partnership as well as about the quantitative facts that can be modified in figures and statistics. Anti-positivism or qualitative analysis methodology relies on the

subjectivist method of dealing with concentrated social events of interest to the scope of research procedures. Hostile to positivism critics scrutinize positivists when they conclude the observations and numbers regarding human actions are not beneficial. Similarly, the Critical Theory Analysis Methodology defines critical and intervention research as research techniques for the investigation of a specific form of research (Dash 2005).

This research is a hybrid approach combining quantitative methods – datamining and qualitative analyses of results obtained from use of machine learning (ML) algorithms.

Figure 1.2 depicts the overview of the methodology used in this research.

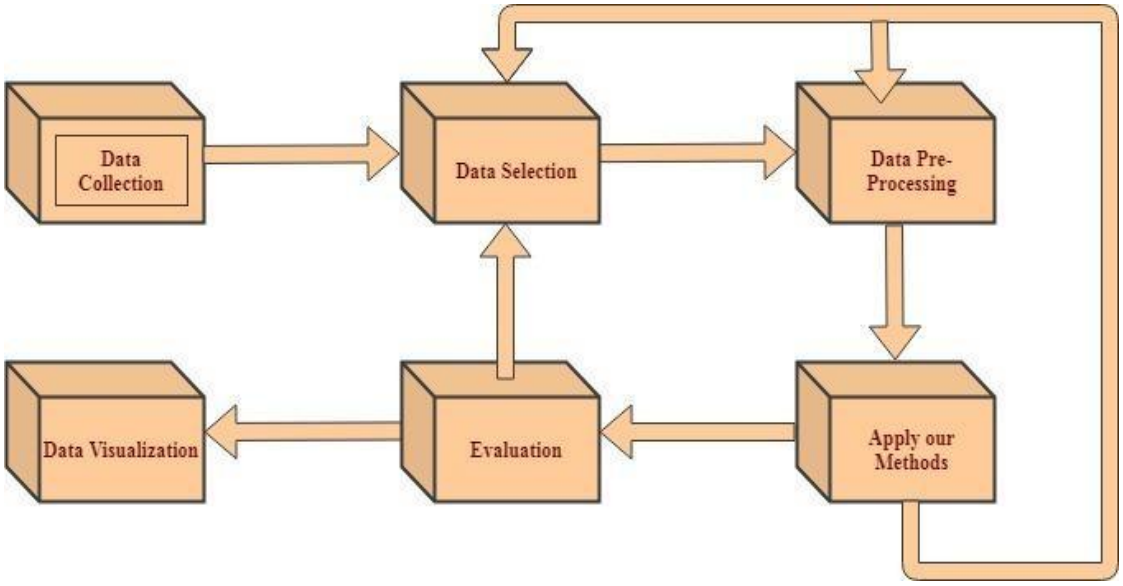


Figure 1.2. Research Method Overview

The steps

1. Data Collection of cardiovascular disease and breast cancer. Some of the data sources used in this thesis are:
   Cardiovascular disease:
   https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
   Breast cancer:
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
   Cervical cancer:

https://archive.ics.uci.edu/ml/machine-learning-databases/00383/
Liver disease:
https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
Diabetes:
https://www.kaggle.com/uciml/pima-indians-diabetes-database
Chronic kidney disease
https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

a) The UCI and KAGGLE machine learning repository are is some of the most popular repositories for machine learning. UCI is an accumulation of databases, field speculations, and generators of information used by analysts in machine learning to prepare and test algorithms for machine learning. David Aha and fellow graduate students at UC Irvine built the store in 1987. Since that time, it has been generally used as an essential source of machine learning databases by students, teachers, and analysts all over the world.

2. Data Processing of "Dirty" data for removal of unwanted data attributes such as deficiency, loudness, and irregularity. The data collection process can establish a dataset with scattered and unreliable information. Inaccurate data have erroneous attribute values; due to data input errors, incorrect data collection instruments, data transmission errors, and customers may only submit inaccurate qualities for compulsory survey fields (Han et al. 2011). For some reason, incomplete data may occur. For instance, during data entry, some attribute values were not necessary and some attribute values were not always available. Irregularity occurs when the record in (Han et al. 2011) dataset conflicts with various records.

3. The completeness, accuracy, and consistency of the data are components that characterize the quality of the data. Pre-processing of data is an essential step towards the process of data mining to fulfill data quality components. Therefore, pre-processing tasks have been carried out in the present research data to ensure that the data set is ready for the mining process to produce as accurate a result as possible. The study proposed a new approach to constructing missing feature values to satisfy the element of completeness; a comparison was also made between the method of selecting features to find the best method suited to datasets

and some techniques for eliminating noise and outliers. At the end of the current phase, the data is ready for the mining process.

4. Data Selection involves in this study, use of feature selection strategies were used for removal of non-significant features, if any,-before starting the mining process. The steps include:

    a. Comparison of different feature selection methods currently used – Wrapper and Filter methods
    b. Introduce Genetic algorithm-based feature selection using logistic regression (LR) as the meta classifier.
    c. Use different techniques as meta classifiers in step b above.
    d. Compare the results obtained from steps a, b and c, and recommend the final method.

5. Data association analysis for exploring the clustering of data and gaining insights. In the literature review, K-means has improved clustering capacity than other unsupervised machine learning methods. K-means analysis was done to see the parameter linkages and deduce the boundary conditions to avoid type 1 and 2 errors.

6. Data mining methods application and evaluation for classification and prediction with highest accuracy.

As far as supervised learning approaches are concerned, logistic regression and decision tree have been identified in most of the literature, with one staying ahead of the other in different research. Support Vector machines or SVM have a place in many research works, but the conclusion of such research excludes its acceptance of other techniques, such as logistic regression and decision tree. Not much has been said about the comparison between the Discriminant analysis and the Logistics regression in terms of accuracy and other aspects. Most of the current research focuses on the combined approach and ensemble methods, namely Random Forest, to ensure the use of the most recent approaches.

The steps include:

a. Application and comparison of performance of ML tools – SVM, Naïve Bayes, Decision Tree, Random Forest, Logistics Regression and Discriminant Analysis to the complete and reduced data set.

An imperative part of the data mining process is the evaluation phase. The purpose of the data mining specialists in this phase is to test and evaluate the model proposed. If the model does not meet the requirements, data mining experts generally rebuild the model at that point by changing its parameters until the desired results are achieved.

In this study, the assessment of the proposed methods is carried out by comparing the results of the model and the genuine data values (class features). The accuracy of the classification and error rate is determined. The error rate (Err) of the classifier is defined as the average number of misclassified samples, divided by the total number of dataset records. On the other hand, it is possible to calculate the classification precision of the model as one minus the rate of error. For example, conventional approaches suggest that if the accuracy of the classification is less than a certain threshold, let us say 80 percent, then some modifications must be made to the method, the feature selection, or the pre-processing phase until satisfactory results are obtained.

The results of different data mining techniques were evaluated, prioritized, and combined in this paper to derive the association rules for predicting critical diseases such as CVD and breast cancer instead of relying on the method of anyone.

Machine Learning Software Development Tools

Five well-known machine learning software has been used for the present research work; SPSS 20, XLSTAT, RapidMiner Studio, Weka 3.8.4, and Python 3.0 programming language. Data Science Experience offers an array of options for a huge amount of data to work with. SPSS consists of some algorithms for data mining to automatically analyze a large collection of data and extract useful knowledge. The XLSTAT add-in for statistical analysis offers a wide range of functions to improve Excel's analytical capabilities, making it the ideal tool for data analysis and statistical requirements (Fahmy & Aubry 1998). To achieve optimum performance and optimum accuracy for the prediction model, the SPSS and XLSTAT parameters have been adjusted. With

RapidMiner Studio, it is possible to access, load, and analyze any form of data, typically structured data and unstructured data such as text, images, and media. It may also collect information and structure unstructured data from these types of data.

Weka is a collection of machine learning algorithms for data mining tasks. This involves pre-processing, grouping, regression, clustering, rules for correlation, and data resources for visualization.

Python is a versatile scripting language that ensures that unlike HTML, CSS, and JavaScript, all forms of programming and software creation, apart from web development, can be used. This covers back-end development, project management, data science, and system scripts, among other aspects.

Results Visualization

Tables, scattered diagrams, bar charts, and figures has been used to display the findings obtained. Data mining professionals determine how to present the data mining findings after the assessment process. Visualization of data helps to encourage the end-user to access and use the findings produced. Since data mining commonly entails retrieving non-existent information from a database, certain concerns about information sources and how to use them may be posed by end-users. In databases, however, end-users require the details to exist on the database already. This thesis has not examined data visualization in-depth. The reasoning is that the present thesis is for research purposes and is not business oriented.

Conclusion

The research methodology used in current research and the source of the dataset used is presented in this chapter. The principal methodologies of the processes of data mining are also described. The research proposes an improved methodology to provide results with good accuracy levels not mentioned in extant studies.

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4585088/

2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4585088/

3. Medication Safety in the Community: A Review of the Literature. 2009 [sighted 2010 01/09/2010]; Available from www.safetyandquality...-con/$File/25953-MS-NPS-LitReview2009.PDF.

4. Article: Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study collected from https://www.eurekalert.org/pub_releases/2019-09/tl-pss083019.php

5. Article: Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study collected from https://www.eurekalert.org/pub_releases/2019-09/tl-pss083019.php

6. ESC Congress 2019 is organised by the European Society of Cardiology and is the world's largest cardiovascular congress with over 500 expert sessions and 11 000 abstracts collected from https://www.eurekalert.org/pub_releases/2019-09/tl-pss083019.php

7. https://www.downtoearth.org.in/news/health/9-6-million-people-will-die-of-cancer-this-year-61646

8. General Information About Male Breast Cancer. 2012 [sighted 2012 30/12/2012]; Available from: http://www.cancer.gov/cancertopics/pdq/treatment/malebreast/Patient/page1.

9. https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d

Chapter 2

Literature Review

Introduction

Today's world poses three problems in the healthcare sector: a shortage of medical practitioners, aging, and increased healthcare expenditure. According to projections by the World Health Organization (WHO), in 2013 global demand and the actual number of health staff were 60.4 million and 43 million, respectively (Scheffler et al. 2016). In 2030 these figures will rise to 81.8 million and 67.3 million. The scarcity of medical facilities continues unsolved and is still severe. From 2000 to 2050 (from 11% to 22%), the global number of adults over 60 years of age will increase (Beard et al. 2016). One primary reason for this is the steady fall in the birth rate and is expected to remain lower over the coming decades. The older the person, the more likely that they would be aged, sick, and require long-term care and prescription services. Consequently, increased human resources and expenditure should be provided to the aging demographic aged 60 or over.

A substantial portion of the gross national product (GDP) is being absorbed as a metric of public health-care spending. WHO report showed that the related figures in China, U.S.A., Canada, Mexico, Russia, India, and Australia were 5.6%, 17.1%, 10.5%, 8.3%, 7.1%, 4.7%, and 9.4% respectively (World Health Organization 2016). These figures will be increasing in the upcoming decades ascribed to aging. It is good news that healthcare services will integrate ML and AI, and thereby becoming smart healthcare, through the increasing rise in computer power and clinical data efficiency. This would certainly help to solve some of the health-care problems listed above. This will likewise aim at fostering sustainable development (Du and Sun 2015; Castro et al. 2015; Momete 2016). The efficiency of care aims to optimize the financial and social value of the health sector efficiently, without compromising the interest of our customers and our ability to deliver coverage in the future. Moreover, inadequate numbers of medical staff and rising proportions of elderly citizens are raising government spending and reducing social performance.

The subject for ML algorithms is about unsupervised learning (Fong et al. 2017; Lim et al. 2017; Sipes et al. 2014; Haque et al. 2015; Kadri et al. 2016; Khan et al. 2017; Ordóñez

et al. 2015; Huang et al. 2015), supervised learning (Nettleton et al. 2010; Akata et al. 2013; Unler et al. 2011; Raducanu and Dornaika 2012; Zhang et al. 2017; Cai et al. 2017; Ichikawa et al. 2016; Muhammad 2015; Wang et al. 2016; Aydın and Kaya Keleş 2017; Li et al. 2016; Li et al. 2016; Miranda et al. 2016; Choi et al. 2017; Hsu et al. 2017; Gu et al. 2014; Chui et al. 2017; Hahne et al. 2014) and semi-supervised learning (Wongchaisuwat et al. 2016; Albalate and Minker 2013; Shi et al. 2015; Reitmaier and Sick 2015; Wang et al. 2015; Nie et al. 2014; Cvetković et al. 2015; Jin et al. 2016; Ashfaq et al. 2017; Yan et al. 2013). This research does not discuss the abstract complexity of these algorithms in terms of technical definition but does include comparisons of accuracies of various DM techniques for revealing potential discrepancies between optimization of ML algorithms and deployment in healthcare. Only two applications in the area of disease management, fatal diseases (cardiovascular disease), and chronic diseases (chronic kidney disease and diabetes) are listed, irrespective of the multitude of smart healthcare innovations. Smart healthcare systems for managing various diseases lead to intelligent decision making. If the performance of the classification algorithm is high in terms of average precision and testing period, it will gradually eliminate the position of medical doctors in diagnosing disease (and medical doctors will still devote their energy to difficult surgery). The second option is that the classification algorithm can be used as a fast test for low cost and large-scale screening (fair overall accuracy and rapid decision).

AI can cause the computer to think. AI's getting machines smarter than ever. The subfield of AI in research is ML. Various scholars claim it is difficult to establish knowledge without understanding. Figure 2.1 demonstrates different facets of the techniques for ML. Types of machine-learning processes include processes of supervised, unsupervised, semi-supervised, Reinforcement, Evolutionary, and deep learning. The data collected is analyzed with certain approaches.

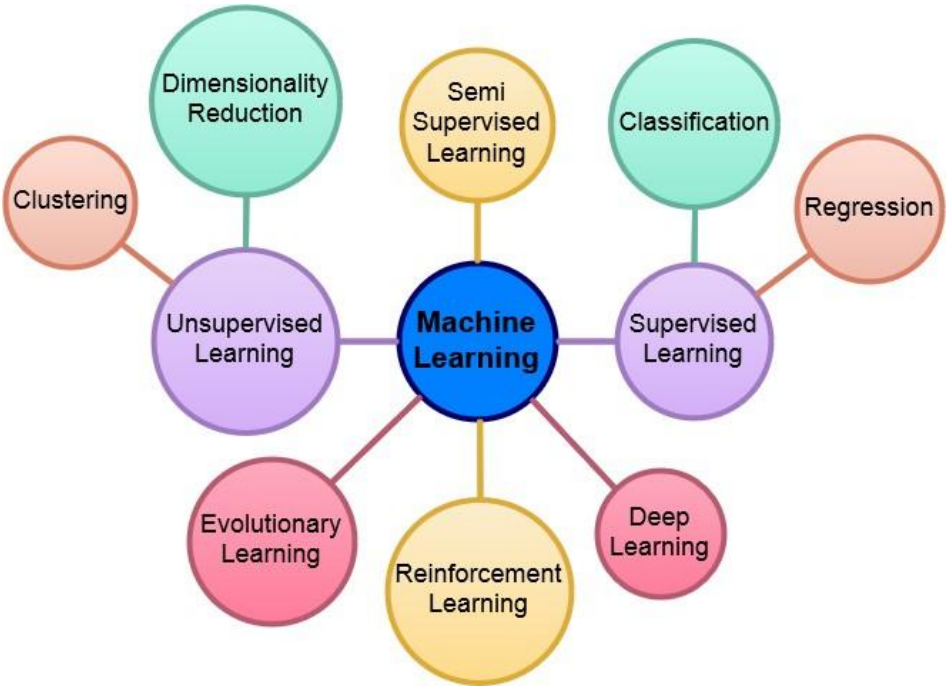


Figure 2.1. Various ML Techniques

Supervised Learning

Supervised learning is when the model is getting trained on a labeled dataset. The labeled dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labeled as shown in the figures below. Given a training set of examples with acceptable goals, and based on this training set, algorithms respond correctly to all feasible inputs. Supervised learning is also known as Learning from exemplars. The Supervised Learning aspects are Classification and regression (Singh and Kumar 2020; Friedman et al. 2001; Hastie et al. 2009).

Classification

By classification algorithms, a class value can be explained or predicted. For many AI applications, classification is an integral part. The classification algorithms are not restricted to two classes and can be used in a variety of categories to classify objects. For instance, it gives a Yes or No prediction, e.g. “Is this malignant tumor?”; “Does this patient have CVD or not?”.

Regression

For training supervised ML, regression techniques are used. Regression methods are generally aimed at describing or predicting a certain numerical value using a previous data set. Regression approaches, for example, will take data from previous instances of heart disease and estimate the probability of heart disease of a patient having similar data. The simplest and most basic approach is known as linear regression.

Unsupervised Learning

Unsupervised learning is the training of a machine using no classified or labeled information that allows the algorithm to handle this information without guidance. Here without previous data preparation, the role of the machine is to group unsorted data according to similarities, trends, and variations. Unsupervised Learning is learning without a guide or teacher. It will work on the datasets automatically and create relationships or patterns on those datasets based on the created relationships. Unsupervised learning methodology tries to assess the variations in input data and classifies the outcomes based on those discrepancies. It is often called size estimation. Clustering needs lessons without supervision (Singh and Kumar 2020; Friedman et al. 2001; Hastie et al. 2009).

Clustering

The role of clustering is to split the population or data points into a variety of groups in such a manner that the data points of the same groups are more similar to the data points of the same group and separate from the data points of other groups. It is simply a list of objects based on their similarity and dissimilarity. It is typically used to find concrete structure, explanatory underlying mechanisms, generative properties, and groupings inherent in a series of examples as a system. It is commonly used in many uses, such as image processing, analysis of data, and identifying patterns.

Semi-Supervised Learning

It is a form of supervised learning technique. Such research also used unlabeled data (usually a lesser amount of labeled data with a hefty amount of unlabeled data) for training purposes. Semi-supervised learning coalesces unsupervised learning (unlabeled data) and supervised learning (labeled data).

Reinforcement Learning

The work has been supported by behavioral psychology. The algorithm has told if the answer is incorrect, it has not indicated how to correct it. It has multiple choices to explore and check before finding the right alternative. It has also been regarded as studying critics. We do not recommend upgrades. Reinforcement learning is distinct from supervised learning in the sense that it does not include appropriate sets of input and output, nor does it specifically define suboptimal behavior. It has also focused on performance on-line.

Evolutionary Learning

This biological evolution analysis has been termed as a method of research: biological organisms have changed to boost their survival rates. Using the fitness description, we will use this principle on a computer (Marshland 2009) to check how accurate the answer is.

Deep Learning

This division of ML is focused on collections of algorithms. These learning algorithms exhibit high-level intellections of outcomes. It uses deep graph with different layers with computation, comprising of both linear and nonlinear transformations.

The commonly used ML algorithms identified among the papers reviewed in this study are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), Artificial Neural Network (ANN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), Deep Neural Network (DNN), Linear Discriminant Analysis (LDA), Multi-Layer Perceptron (MLP), Genetic Algorithm (GA), Neural Network (NN), Radial Basis Function (RBF), Sequential Minimal Optimization (SMO), Ant Colony Optimization (ACO), Adaptive Boosting (AdaBoost), Gradient Boosting (GB). Data repositories utilized for research purposes are UCI Machine Learning Dataset Repository (Heart Disease - UCIMLR, Cleveland Heart - UCI-A, Hungarian Heart - UCI-B, Statlog Heart - UCI-C, SPECTF - UCI-D, Long Beach Heart - UCI-E, Switzerland Heart - UCI-F, Apollo Hospitals CKD Dataset - CKD-A, Pima Indians Diabetes - PID) and Framingham Heart Study Dataset (FHS). In Table 2.1., differences among commonly used ML algorithms (Atlas et al. 1990; Uddin et al. 2019) are tabulated.

Table 2.1. Differences Among Various Supervised ML Algorithms

Supervised ML Algorithm	Advantages	Drawbacks
SVM	• It is much more stable than LR. • Able to manage several feature spaces. • The chance of over-fitting is very little. • Classification accuracy for semi-structured or unstructured files, such as texts, images, etc. is very high.	• Computational cost is very high for massive and complicated datasets. • In the case of noisy data, performance is very poor. • The output model, weight, and effect of features are often difficult to interpret. • Generic SVM is unable to classify more than two classes until it has been expanded.
LR	• Simple to implement and easy to execute. • Models based on LRs can be modified quickly. • Does not make any predictions about the propagation of independent feature(s). • It's got a good probabilistic description of the model parameters.	• Not able to score a reasonable precision in cases where input variables have complex relationships. • Does not take into account the causal relationship between the features. • LR's main components-logic structures are susceptible to overconfidence. • Can overstate the accuracy of the forecast due to the bias of the sampling. • Generic LR can only distinguish features with two states (i.e., dichotomous) except they are multinomial.
KNN	• An easy algorithm which can easily distinguish the instances. • Can manage noisy instances or instances with incorrect, incomplete or, missing attribute values. • Able to perform both classification and regression operations.	• Algorithmically costly when the number of attributes is increased. • Attributes are assigned equivalent weight, which can lead to poor classification results. • Provide no details on which features are most important for successful classification.
DT	• The resulting classification tree is simpler to grasp and analyze.	• Include classes that are mutually exclusive.

	• The pre-processing of data is much easier than other techniques. • Various data types, such as numerical, nominal, categorical, are equally accepted. • It can produce robust classifiers using ensemble techniques and can be checked by means of statistical testing.	• The algorithm cannot split if a non-leaf node attribute or feature value is absent. • The algorithm is based on the order of the attributes or features. • May not do as good as every other classifier (e.g., the Artificial Neural Network).
NB	• Easy and very helpful to classify massive datasets • Equally effective on both binary and multi-class classifications. • It needs fewer data to train the model. • It can render probabilistic forecasts, and can also manage both continuous and discreet values in the datasets.	• Classes ought to be mutually exclusive in nature. • The existence of a correlation between attributes adversely affects the efficiency of the classification. • The standard distribution of quantitative attributes is presumed.
ANN	• It can identify dynamic non-linear interactions between dependent and independent variables. • Needs less rigorous mathematical and statistical knowledge. • Existence of several training algorithms. • Can be extended to issues of classification and regression.	• The client does not have access to the exact decision-making mechanism and thus it is computationally costly to train the network for a complex classification task. • Independent features must be pre-processed.

Cardio Vascular Disease (CVD)

The next segment of this research discusses the various DM approaches used to forecast heart disease. WHO reckons that 12 million deaths occur due to heart ailment globally every year. 17.9 million people die of heart disease each year. Around 31% of the world's deaths come from heart disease. WHO expects that by 2030, nearly 23.6 million individuals will lose their lives from heart disease (Dangare and Apte 2012). The overwhelming amount of data generated to predict heart disease is too complicated and voluminous to be interpreted and analyzed using conventional methods. Using data processing tools, it takes less time to forecast the disease more accurately. In the following section, different papers in which one or more DM algorithms have been used to predict heart disease have been studied. Applying DM techniques to cardiac disease treatment data can provide the same reliable performance as achieved in cardiac disease diagnosis. This chapter analyzes various data analysis methods employed by scientific researchers or clinicians for a successful diagnosis of heart disease as demonstrated in Table 2.2., Table 2.6., and Figure 2.4..

Table 2.2. Comprehensive Analysis of ML Methods for Diagnosing Heart Disease

Author & Year	Data Source	ML Technique	Accuracy (%)
(Ramani et al. 2020)	UCI-B	KNN	71.18
		CART	81.35
		SVM	62.71
		GNB	84.75
		ANN	91.45
(Panda and Dash 2020)	UCI-A	GNB	91.66
		RF	85.00
		DT	76.66
		LR	86.66
		Extra Trees (ensemble)	86.66
		KNN	70.00
		SVM	90.00
(Pathak and Valan 2020)	UCIMLR	Fuzzy set with DT	88
		Framingham Risk Score (FRS)	52.33

(Amarbayasgalan et al. 2020)	Korea National Health and Nutrition Examination Survey Dataset	NB	73.06
		KNN	78.76
		DT	75.13
		RF	81.11
		SVM	80.45
		DNN	82.67
(Sharma et al. 2020)	UCI-C	Features Extraction using Binary ACO, FA, PSO and Artificial Bee Colony (ABC) along with ML classifiers DT, KNN, SVM and RF	Max Accuracy for DT using BPSO = 90.093
(Magesh and Swarnalatha 2020)	UCI-A	Cluster based DT Learning (CDTL) + RF	89.30
		RF	85.90
		SVM	67
		LM	61.90
		DT	60.20
(Mienye et al. 2020)	Kaggle FHS	ANN	85
		KNN	81
		CART	76
		LR	83
		NB	82
		LDA	83
		Sparse Auto Encoder (SAE) + ANN	90
(Ayon et al. 2020)	UCI-C	SVM	97.41
		LR	96.29
		DNN	98.15
		DT	96.42
		NB	91.38
		RF	90.46
		KNN	96.42
	UCI-A	SVM	97.36
		LR	92.41
		DNN	94.39
		DT	92.76
		NB	91.18

		RF		89.41
		KNN		94.28
(Khourdifi and Bahaj 2019)	UCI-A	Classifiers optimized by Fast Correlation-Based Feature Selection (FCBF), PSO and ACO	KNN	99.65
			SVM	83.55
			RF	99.6
			NB	86.15
			MLP	91.65
(Akgül et al. 2019)	UCI-A	ANN		85.02
		Hybrid approach combining ANN and GA		95.82
(Krishnani et al. 2019)	FHS	RF		96.80
		DT		92.45
		KNN		92.81
(Desai et al. 2019)	UCI-A	Back-Propagation NN		85.07
		LR		92.58
(Ali et al. 2019)	UCI-A	X^2 statistical model + DNN		91.57 (k-fold)
				93.33 (holdout)
(Amin et al. 2019)	UCI-C	Vote		87.41
		NB		84.81
		SVM		85.19
(Bashir et al. 2019)	UCIMLR	DT		82.22
		LR		82.56
		RF		84.17
		NB		84.24
		LR (SVM)		84.85
(Burse et al. 2019)	UCI-A	Multi-Layer Pi-Sigma Neuron Network (MLPSNN) + Normalization		91.34
		MLPSNN + PCA		94.53
		MLPSNN + k-fold		90.44
		SVM-LDA		88.32
(Saqlain et al. 2019)	UCI-A	Reverse feature selection algorithm (RFSA), Feature subset selection algorithm (FSSA),		81.19
	UCI-B			84.52
	UCI-F			92.68

	UCI-D	Forward feature selection algorithm (FFSA), Mean Fisher score-based feature selection algorithm (MFSFSA), RBF kernel-based SVM	82.70
(Ottom and Alshorman 2019)	UCI-A	NB	83.5
		SVM	84.2
		KNN	78.9
		ANN	81.9
		J48	79.2
(Kannan and Vasanthi 2019)	UCI-A	LR	86.5
		RF	80.9
		Stochastic Gradient Boosting (SGB)	84.3
		SVM	79.8
(Paul et al. 2018)	UCI-A	Adaptive fuzzy decision support system using GA and Modified dynamic multi-swarm particle swarm optimization (MDMS-PSO)	92.31
	UCI-B		95.56
	UCI-F		89.47
(Rajliwall et al. 2018)	NHANES Physical Activity and CVD Fitness Data from National Centre for Health Statistics in Centre for Disease Control	NB	95.7
		Bagging	96.5
		DT	97.6
		LR	96.4
		KNN	80.8
		RF	98.5
		SVM	95.4
		NN	98.8
	FHS	KNN	90.1
		RF	90.1
		SVM	90.2
		NB	89.9
		DT	90
		LR	90
		Ensemble	89.3
		NN	89
(Dwivedi 2018)	UCI-C	NB	83
		Classification Tree	77
		KNN	80

		LR	85
		SVM	82
		ANN	84
(Shylaja and Muralidharan 2018)	UCI-A	ANN	85.30
		NB	81.14
		Repeated Incremental Pruning to Produce Error Reduction (RIPPER)	81.08
		C4.5 DT	79.05
		SVM	85.97
		KNN	84.12
(Sabahi 2018)	Coronary Heart Disease Data from Iranian Hospitals	Bimodal Fuzzy Analytic Hierarchy Process (BFAHP)	85.91 (mean)
	UCI-B		86.57
	UCI-A		87.31
	UCI-E		85.62
	UCI-F		85.22
(Kurian and Lakshmi 2018)	UCI-A	KNN	81.2
		DT	79.06
		NB	86.4
		Ensemble Classifier (KNN, DT, NB)	90.8
(Rahman et al. 2018)	UCI-A	NB	89.32
		LR	84.47
		NN	90.2
		Ensemble of NB, LR & NN	91.26
(Uyar and Ilhan 2017)	UCI-A	GA based trained Recurrent Fuzzy Neural Networks (RFNN)	Training Set = 96.43, Testing Set = 97.78, Overall = 96.63
(Liu et al. 2017)	UCI-C	ReliefF and Rough Set (RFRS), Heuristic Rough Set Reduction, C4.5 DT, Jackknife Cross Validation	92.59

(Arabasadi et al. 2017)	Z-Alizadeh Sani dataset	NN + GA	93.85
	UCI-B		87.1
	UCI-A		89.4
	UCI-E		78.0
	UCI-F		76.4
(Vivekanandan and Iyengar 2017)	UCI-A	Modified differential evolution (DE) algorithm with Integrated model of fuzzy analytic hierarchy process (AHP) and feed-forward neural network	83
(Pouriyeh et al. 2017)	UCI-A	DT	77.55
		NB	83.49
		KNN	83.16
		MLP	82.83
		RBF	83.82
		Single Conjunctive Rule Learner (SCRL)	69.96
		SVM	84.15
(Zriqat et al. 2016)	UCI-A	DT	99.01
		NB	78.88
		Discriminant	83.50
		RF	93.40
		SVM	76.57
	UCI-C	DT	98.15
		NB	80.37
		Discriminant	82.59
		RF	91.48
		SVM	75.56
(Sultana et al. 2016)	UCIMLR	KStar	75.19
		J48	76.67
		SMO	84.07
		Bayes Net	81.11
		MLP	77.41
	Patients' medical data from Enam Medical Diagnosis	KStar	75
		J48	86
		SMO	89
		Bayes Net	87

	Centre, Savar, Dhaka, Bangladesh	MLP		86
(Long et al. 2015)	UCIMLR	Rough Sets based Attribute Reduction and Interval Type-2 Fuzzy Logic System (IT2FLS)	Binary Particle Swarm Optimization and Rough Sets based Attribute Reduction (BPSORS-AR)	87.0
			Chaos Firefly Algorithm and Rough Sets based Attribute Reduction (CFARS-AR)	88.3
	UCI-D	IT2FLS	BPSORS-AR	81.8
			CFARS-AR	87.2

2.3. Chronic Kidney Disease (CKD)

Chronic kidney disease (CKD) is an intensifying problem for well-being and one of the foremost causes of death worldwide (Nahas 2005; Alebiosu and Ayodele 2005). It is a secret, highly complex, and progressive condition that interferes with the physiological functions of some organs, including the cardiovascular system (Mahdavi-Mazdeh 2010; Khalkhaali et al. 2010). CKD experiences high health care costs. Most nations have now based a great deal on early diagnosis and disease prevention. The World Health Organization has estimated the cost of dialysis at $1.1 trillion in the past decade (Lysaght 2002). Due to the intricate nature of the renal disease, its early covert existence, and the heterogeneity of patients, it is difficult to predict the timing of worsening kidney disease with high accuracy (Fauci 2008). If CKD is diagnosed early, much of its signs can be averted or at least postponed. Progression of renal failure can see as a result of multiple causes, including intrinsic renal dysfunction, obesity, proteinuria, age, and many others (Mahdavi-Mazdeh et al. 2012; Indridason et al. 2007).

The dynamic aspect of CKD, its hidden existence at an early level, and the variability of patients need a reliable estimation of the deteriorating timeline of renal function. CKD is a dynamic disorder with strong non-linear nature and the specific variables affect the extent and degree with the development of the disease (Nahas 2005; Mahdavi-Mazdeh 2010; Khalkhaali et al. 2010). A robust statistical model for CKD would also use a consistent index. GFR is the most effective metric for assessing renal activity and physiological status (Mahdavi-Mazdeh et al. 2012; Indridason et al. 2007; Kumar et al. 2013). If GFR discrepancies in the CKD patient may be determined, the 15 cc/min/1.73 m^2 time for renal replacement therapy has more reliably expected and the surgical care protocols improved. Progression of renal disease can see as a result of several causes, including intrinsic renal dysfunction, weight, proteinuria, age, among several others. Various studies and research groups have attempted to measure and forecast variations in GFR using observational formulae based on comparative analysis of data attained from patients with CKD (Walser 1994; Walser et al. 1993). This chapter analyzes various ML methods employed by scientific researchers or clinicians for a successful diagnosis of CKD as demonstrated in Table 2.3., Table 2.7., and Figure 2.5..

Table 2.3. Comprehensive Analysis of ML Methods for Diagnosing Chronic Kidney Disease

Author & Year	Data Source	ML Technique		Accuracy (%)
(Rubini and Perumal 2020)	CKD-A	Fruit Fly Optimization Algorithm for Feature Selection, Classification using Multi-Kernel Support Vector Machine (MKSVM)		98.51
	UCI-A			96.04
	UCI-B			93.197
	UCI-F			95.12
(Alloghani et al. 2020)	Ambulatory Electronic Medical Record (EMR) of Patients from Tawam Hospital, Al Ain, UAE	KNN		91.3
		DT		88.5
		RBF SVM		91.1
		Polynomial SVM		91.7
		Ridge LR		90.0
		LASSO LR		90.4
		Stochastic Gradient Descent NN		90.4
		Logistic NN		90.4
		NB		83.0
		CN2 Induction Rule		85.7
		RF		90.2
		Boosted DT		88.5
(Sobrinho et al. 2020)	CKD Dataset from University Hospital located at the Federal University of Alagoas (UFAL), Brazil.	J48	Correctly Classified Instances	95.00
		RF		93.33
		NB		88.33
		SVM		76.66
		MLP		75.00
		KNN		71.67
(Alaiad et al. 2020)	CKD-A	NB		94.50
		DT (J48)		96.75
		SVM		97.75
		KNN		98.50
		Association Rule (JRip)		96.00
(Elhoseny et al. 2019)	CKD-A	Density based Feature Selection (DFS) with ACO		95.00
	CKD-A	LOG (Regression-based Model)		98.95

(Qin et al. 2019)		RF (Tree-based Model)	99.75
		Integrated Model of LOG & RF	99.83
(Rabby et al. 2019)	CKD-A	KNN	71.25
		SVM	97.50
		RF	98.75
		GNB	100
		AdaBoost Classifier	98.75
		LDA	97.50
		LR	97.50
		DT	100
		GB	98.75
		ANN	65
(Hasan and Hasan 2019)	CKD-A	AdaBoost	99
		Bootstrap Aggregating	96
		Extra Trees	98
		GB	97
		RF	95
(Besra and Majhi 2019)	CKD-A	NB	99.10
		SMO	99.55
		IB1	99.90
		Multiclassifier	98.30
		VFI	99.48
		RF	99.43
(Almansour et al. 2019)	CKD-A	ANN	99.75
		SVM	97.75
(Saringat et al. 2019)	CKD-A	ZeroR	62.50
		Rule Induction	92.50
		SVM	90.25
		NB	98.50
		DT	95.50
		Decision Stump	92.00
		KNN	94.75
		Classification via Regression	98.25
(Almasoud and Ward 2019)	CKD-A	LR	98.75
		SVM	97.5
		RF	98.5

			GB	99.0
(Saha et al. 2019)	Chronic Kidney Disease Dataset from National Kidney Foundation, Bangladesh		RF	96.75
			NB	95.04
			MLP	97.03
			LR	95.75
			Adam DNN	97.34
(Pasadana et al. 2019)	CKD-A		Decision Stump	92
			Hoeffding Tree	95.75
			J48	99
			CTC	97
			J48graft	98.75
			LMT	98
			NB Tree	98.5
			RF	100
			Random Tree	95.5
			REPTree	96.75
			Simple CART	97.5
(Ogunleye and Qing-Guo 2019)	CKD-A		Recursive Feature Elimination (RFE)	98.9
			Extra Tree Classifier (ETC)	97.9
			Univariate Selection (US)	97.9
			Collaborating RFE, ETC & US with optimal XGBoost modelling	100.0
(Kriplani et al. 2019)	CKD-A		NB	97.77
			DNN	97.77
			LR	97.32
			RF	99.11
			AdaBoost	98.21
			SVM	98.21
(Ripon 2019)	Out of 2800 patients 1050 are healthy and 1750 adults have Chronic Kidney Disease	Classification	AdaBoost	99
			LogitBoost	99.75
		Decision Rules	J48	99
			Ant-Miner (ACO)	99.5

	[Source Unknown]			
(Devika et al. 2019)	CKD-A		NB	99.635
			KNN	87.78
			RF	99.844
(Alaoui et al. 2018)	CKD-A		XGBoost Linear	100
			XGBoost Tree	99.75
			Linear SVM	99.5
			C&R Tree	99.25
			CHAID	98.5
			Quest	98.25
			C5	98
			Random Trees	96.528
			Tree – AS	90.5
			Discriminant	88.5
(Zeynu and Patil 2018)	CKD-A	Feature Selection using Info Gain Attribute Evaluator with Ranker Search or, using Wrapper Subset Evaluator with Best First Search	KNN	99
			J48	98.75
			ANN	99.5
			NB	99
			SVM	98.25
			Ensemble Model by combing five heterogeneous classifiers based on voting algorithm	99
(Kemal 2018)	CKD-A	Feature Selection by PSO	KNN	95.75
			SVM	98.25
			RBF	98.75
			Random Subspace	99.75
(Alassaf et al. 2018)	Saudi CKD Dataset retrieved from King Fahd University Hospital (KFUH), Khobar		ANN	98
			SVM	98
			NB	98
			KNN	93.9

(Aljaaf et al. 2018)	CKD-A	Classification and Regression Tree (RPART)	95.6
		SVM	95.0
		LR	98.1
		MLP	98.1
(Tikariha and Richhariya 2018)	CKD-A	KNN	98.5
		SVM	97.75
		NB	94.5
		C4.5	96.75
(Hore et al. 2018)	CKD-A	NN	98.33
		RF	92.54
		MLP-FFN	99.5
		NN-GA	100
(Ahmad et al. 2017)	CKD-A	SVM	98.34
(Alasker et al. 2017)	CKD-A	ANN	98.41
		NB	99.37
		Decision Table	97.62
		J48	98.41
		One Rule Decision Tree	97.62
		KNN	99.21
(Borisagar et al. 2017)	CKD-A	Levenberg Marquardt Back Propagation	99.8
		Bayesian Regularization Back Propagation	99.5
		Scaled Conjugate Gradient Back Propagation	98.7
		Resilient Back Propagation	99.5
(Chatterjee et al. 2017)	CKD-A	Multilayer Perceptron Feedforward Network (MLP-FFN)	96.33
		Genetic Algorithm trained Neural Network (NN-GA)	97.5
		Cuckoo Search trained Neural Network (NN-CS)	99.2
(Gunarathne et al. 2017)	CKD-A	Multiclass Decision Forest	99.1
		Multiclass Decision Jungle	96.6
		Multiclass LR	95.0
		Multiclass NN	97.5

(Polat et al. 2017)	CKD-A	SVM	Classifier Subset Evaluator with Greedy Stepwise Search Engine		98
			Wrapper Subset Evaluator with Best First Search Engine		98.25
			Correlation Feature Selection Subset Evaluator with Greedy Stepwise Search Engine		98.25
			Filtered Subset Evaluator with Best First Search Engine		98.5
(Sisodia and Verma 2017)	CKD-A		NB		97
			SMO		99
			J48 DT		99
			RF		100
			Bagging		98
			AdaBoost		99
(Subasi et al. 2017)	CKD-A		ANN		98
			SVM		98.5
			KNN		95.75
			C4.5		99
			RF		100
(Wibawa et al. 2017)	CKD-A	Correlation-based Feature Selection (CFS), AdaBoost	NB		98
			KNN		98.1
			SVM		97.5
(Basar and Akan 2017)	CKD-A	REPTree	Adaboost	Information Gain Attribute Evaluator Feature Selection	99
			Bagging		98.5
			RF		99.75
		Best First Decision Tree (BFTree)	Adaboost		99.5
			Bagging		99.75
			RF		100
		J48 DT	Adaboost		99.5
			Bagging		99.75
			RF		99.25
		SVM	Adaboost		98.5

		Bagging	97.75
		RF	97.5
(Tazin et al. 2016)	CKD-A	NB with Ranking Algorithm (RA)	96
		SVM with RA	98.5
		DT with RA	99
		KNN with RA	97.5
(Charleonnan et al. 2016)	CKD-A	SVM with Gaussian kernel	98.3
		LR	96.55
		DT	94.8
		KNN	98.1
(Chen et al. 2016)	CKD-A	Fuzzy Rule-building Expert System (FuRES)	99.6
		Fuzzy Optimal Associative Memory (FOAM)	98
		Partial Least Squares Discriminant Analysis (PLS-DA)	95.5
(Chetty et al. 2015)	CKD-A	WrapperSusetEval Attribute Evaluator with NB classifier and Best First Search	99
		WrapperSusetEval Attribute Evaluator with SMO classifier and Best First Search	98.25
		WrapperSusetEval Attribute Evaluator with IBK classifier and Best First Search	100

Diabetes Mellitus

Diabetes mellitus universally identified as diabetes is a chronic condition and one of the world's primary metabolic diseases (Kandhasamy and Balamurali 2015; Kang et al. 2015). It is associated with an abnormal increase in blood glucose (hyperglycemia) due either to inadequate pancreatic insulin production (Type 1 diabetes) or to the failure of cells to respond effectively to pancreatic insulin (Type 2 diabetes) (Vijiyarani and Sudha 2013). The danger to all this instability in plasma glucose (hyperglycemia, hypoglycemia) is that there is substantial damage to several vital systems of the body, especially the blood vessels and nervous system (Lukmanto and Irwansyah 2015). While its origins are not yet fully known, scientists agree that both genetic causes and environmental factors are concerned (Barakat et al. 2010). Nevertheless, diabetes was prevalent in the severe form mostly in adults, and formerly known to be 'adult-onset' diabetes. It is now broadly agreed that diabetes mellitus is closely related to the effects of aging.

Despite the above-mentioned consequences, early diagnosis and monitoring of diabetes are the standard prerequisites. Electronic Medical Records (EMRs) play an essential role in this field by keeping track of regular safety tests that are important to the patient's condition over time. Diabetes risk management models and their various algorithms have been extensively researched in order to provide a fast and comprehensive analysis of scientific evidence. Schwarz et al. (2009) offered a thorough review of these models in terms of accuracy and sensitivity. However, because such risk-scoring models entail human intervention, although to some degree in the calculation of criteria and risk-scoring, the outcomes could be susceptible to human error. DM is a crucial tool for science repositories. This revolutionary approach improves the sensitivity and/or specificity of disease detection and diagnosis by providing a comparatively enhanced variety of methods. This also greatly decreases indirect costs by reducing repetitive and expensive laboratory procedures (Canlas 2009). Extensive diabetes prediction experiments have been undertaken over a range of years. Several studies have recently compared various approaches to learning. Such analyses are usually rare and are performed on a small range of data sets available from Pima Indian Diabetic Database. This chapter analyzes various data analysis methods employed by scientific researchers or clinicians for a successful diagnosis of Diabetes as demonstrated in Table 2.4., Table 2.8., and Figure 2.6..

Table 2.4. Comprehensive Analysis of ML Methods for Diagnosing Diabetes Mellitus

Author & Year	Data Source	ML Technique		Accuracy (%)
(Singh and Singh 2020)	PID	Non-dominated Sorting Genetic Algorithm II (NSGA-II) Stacking with Multi-Objective Optimization		83.80
(Choubey et al. 2020)	PID	LR		78.6957
		KNN		73.4783
		ID3 DT		75.6522
		C4.5 DT		76.5217
		NB		76.96
		LR	Principal Component Analysis (PCA)	79.5652
		KNN		73.913
		ID3 DT		75.6522
		C4.5 DT		74.7826
		NB		78.6957
		LR	PSO	79.5652
		KNN		73.913
		ID3 DT		75.6522
		C4.5 DT		74.7826
		NB		78.69
	Localized Diabetes Dataset from Bombay Medical Hall, Mahabir Chowk, Pyada Toli, Upper Bazar, Ranchi, Jharkhand, India	LR		92.429
		KNN		85.1735
		ID3 DT		81.388
		C4.5 DT		95.5836
		NB		92.11
		LR	PCA	93.6909
		KNN		85.8044
		ID3 DT		81.388
		C4.5 DT		95.5836
		NB		92.1136
		LR	PSO	93.3754
		KNN		88.6435
		ID3 DT		81.388
		C4.5 DT		94.0063
		NB		92.43

(Maniruzzaman et al. 2020)	Diabetes [2009–2012] Dataset from National Health and Nutrition Examination Survey (NHANES)	Feature Selection using LR	NB	2 fold cross-validation protocol (K2)	86.42
				5-fold cross-validation protocol (K5)	86.61
				10-fold cross-validation protocol (K10)	86.70
			DT	K2	89.90
				K5	89.97
				K10	89.65
			Adaboost	K2	91.32
				K5	92.72
				K10	92.93
			RF	K2	93.12
				K5	94.15
				K10	94.25
(Shuja et al. 2020)	Primary clinical dataset containing record of 734 patients from a diagnostic lab in Kashmir valley	Bagging		Without SMOTE (Synthetic Minority Oversampling Technique)	94.1417
				With SMOTE	94.2197
		SVM		Without SMOTE	90.7357
				With SMOTE	89.0173
		MLP		Without SMOTE	93.4605
				With SMOTE	93.8348
		LR		Without SMOTE	92.2343
				With SMOTE	90.2697

		DT	Without SMOTE	92.5068
			With SMOTE	94.7013
(Devi et al. 2020)	PID	Farthest First (FF) Clustering Algorithm with SMO Classifier		99.4
(Yuvaraj and SriPreethaa 2019)	Pima Indians Diabetes Database from National Institute of Diabetes and Digestive Diseases	DT		88
		NB		91
		RF		94
(Singh and Gupta 2019)	PID	Fuzzy Rule Miner ANT-FDCSM (Fuzzy Distinct Class based Split Measure)		87.7
(Choudhury and Gupta 2019)	PID	SVM		75.68
		KNN		75.1
		DT		67.57
		NB		76.64
		LR		77.61
(Kumar and Manjula 2019)	PID	8-12-8-1 MLP (no. of nodes in input layer – hidden layer – hidden layer – output layer)		79.05
		8-32-32-1 MLP		83.02
		8-64-64-1 MLP		84.26
		8-128-128-1 MLP		86.67
(Jayashree and Kumar 2019)	PID	Back Propagation Neural Network Optimized with Cuckoo Search Algorithm (BPNN-CSA)		94.37
		Learning Vector Quantization Optimized with Ant Colony (LVQAC)		95.841
		Memetic Optimized Deep Learning Neural Network (MODLNN)		97.27
		Hybrid Swarm Intelligent Redundancy Relevance (SIRR) with Convolution Trained Compositional Pattern Neural Network (CTCPNN)		99.35
(Gbengaa et al. 2019)	PID	NB		76.3021
		RF		73.758

		J48	73.8281
		MLP	73.3906
		Random Tree	68.099
		Modified J48	99.8701
		Non-Nested Generalisation Exemplars (NNGE) Classifier	100
(Xie et al. 2019)	2014 BRFSS (Behavioral Risk Factor Surveillance System, 2014) Data from CDC	NN	82.41
		LR	80.68
		Linear SVM	80.82
		RBF SVM	81.78
		RF	79.27
		NB	77.56
		Polynomial SVM	79.62
		DT	74.26
(Pei et al. 2019)	Annual physical examination reports from electronic health records database in Shengjing Hospital of China Medical University	AdboostM1	91.27
		J48	95.03
		SMO	90.78
		NB	89.34
		Bayes Net	88.78
(Birjais et al. 2019)	PID	GB	86
		LR	79.2
		NB	77
(Vigneswari et al. 2019)	PID	RF	78.54
		C4.5	76.25
		Random Tree	72.41
		REPTree	75.48
		Logistic Model Tree (LMT)	79.31
(Xiong et al. 2019)	Physical Examination Data (Diabetic & Non-Diabetic Patients) from Nanjing Drum Tower Hospital, China	MLP	87
		AdaBoost	86
		Trees Random Forest (TRF)	86
		SVM	86
		Gradient Tree Boosting (GTB)	86

(Xu and Wang 2019)	PID	Weighted Feature Selection Algorithm based on Random Forest (RF-WFS) + XGBoost classifier		93.75
(Giveki and Rastegar 2019)	PID	Rough Set Theory (RST) + Radial Basis Function Neural Networks - Harmony Search (RBFNN-HS)	70% train – 30% test	96.5
			90% train – 10% test	98.7
(Rawat and Suryakant 2019)	PID	AdaBoost		79.68
		LogitBoost		78.64
		RobustBoost		78.64
		NB		76.04
		Bagging		81.77
(Bani-Hani et al. 2019)	PID	SVM with GA		79.72
		MLP with GA		76.88
		RF with Grid Search (GS)		77.50
		Probabilistic Neural Network (PNN)		71.03
		GNB		79.09
		KNN with GS		76.59
		General Regression Neural Network Oracle (GRNN O.)		79.54
		Recursive General Regression Neural Network Oracle (R. GRNN O.)		81.14
(Abed and Ibrikci 2019)	PID	Gradient Descent		76.3
		Gradient Descent with Variable Learning Rate		78.6
		Levenberg-Marquardt Back Propagation		79.9
		BFGS Quasi-Newton Back Propagation		77.1
		MLP Bayesian Regularization		96
		Classification Naive Bayes (CNB)		78.8
		SVM (RBF Kernel)		76.6
		KNN		76.6

		LDA	80.5
(Manikandan 2019)	PID	ACO with Fuzzy Rule	71.4285
		Grey Wolf Optimization (GWO) with Fuzzy Rule	81.1585
(Krijestorac et al. 2019)	PID	LR	77.90
		DT (Medium)	77.10
		Linear SVM	77.40
		Cosine KNN	76.40
(Kaur and Kumari 2018)	PID	Linear Kernel Support Vector Machine (SVM-linear)	89
		Radial Basis Function Kernel Support Vector Machine (RBF-SVM)	84
		KNN	88
		ANN	86
		Multifactor Dimensionality Reduction (MDR)	83
(Sisodia and Sisodia 2018)	PID	NB	76.30
		SVM	65.10
		DT	73.82
(Wu et al. 2018)	PID	Improved K-means Cluster Algorithm & LR	95.42
(Wei et al. 2018)	PID	LR with Parameter Optimization	77.47
		DNN	77.86
		SVM	77.60
		DT	76.30
		NB	75.79
(Mir and Dhage 2018)	PID	NB	77
		SVM	79.13
		RF	76.5
		Simple CART	76.5
(Husain and Khan 2018)	NHANES 2013-14 Diabetes Dataset	LR	95.5
		KNN	96.1
		GB	96
		RF	96
		Ensemble Method using Majority Voting	96
	PID	SVM	78.05

(Dey et al. 2018)		KNN	Min Max Scaler (MMS) Normalization	75.5
		NB		79.3
		ANN		82.35
(Dwivedi 2018a)	PID	ANN		77
		SVM		74
		KNN		73
		NB		75
		LR		78
		Classification Tree		70
(Vijayan & Anjali 2015)	PID for Training & Testing, Local Dataset from different places of Kerala for Validation	DT	AdaBoost	77.6
		SVM		79.687
		NB		79.687
		Decision Stump		80.72
(Nilashi et al. 2017)	PID	PCA + Self Organizing Map (SOM) Clustering + NN		92.28
(Bhatia and Syal 2017)	PID	K-Means + GA + SVM		95.946
		K*-Means + GA + SVM		97.959
(Chen et al. 2017)	PID	K-means Clustering with J48 DT Classifier		90.04
(Erkaymaz et al. 2017)	PID	Newman–Watts Small-World FeedForward Artificial Neural Network (SW FFANN)		93.06
(Choubey et al. 2017)	PID	NB		76.9565
		GA + NB		78.6957
(Komi et al. 2017)	Unknown	Gaussian Mixture Model (GMM) Classifier		81
		ANN		89
		Extreme Learning Machine (ELM)		82
		LR		64
		SVM		74
(Jahangir et al. 2017)	PID	Automatic Multilayer Perceptron (AutoMLP) with Enhanced Class Outlier Detection using Distance Based Algorithm		88.7
(Maniruzzaman et al. 2017)	PID	LDA	RBF Kernel with K10 cross-validation protocol	77.86
		Quadratic Discriminant Analysis (QDA)		76.56

		NB		77.57
		Gaussian Process Classification (GPC)		81.97
(Hayashi and Yukita 2016)	PID	Sampling Re-RX with J48graft		83.83
(Panwar et al. 2016)	PID	KNN + PCA		100
(Kamadi et al. 2016)	PID	PCA for Dimensionality Reduction, Modified Gini Index based Fuzzy SLIQ Decision Tree algorithm to construct Decision Rules		76.8
(Kandhasamy and Balamurali 2015)	PID	J48	without noisy data (after pre-processing)	86.46
		KNN		100
		SVM		77.73
		RF		100
(Nai-arun and Moungmai 2015)	Information of 30122 people collected from 26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital, Thailand	DT		85.090
		ANN		84.532
		LR		82.308
		NB		81.010
		Bagging with DT		85.333
		Bagging with ANN		85.324
		Bagging with LR		82.318
		Bagging with NB		80.960
		Boosting with DT		84.098
		Boosting with ANN		84.815
		Boosting with LR		82.312
		Boosting with NB		81.019
		RF		85.558

Breast cancer

Breast cancer is the most severe malignancy in the women population. When stratified by age, 1 in 24 women develops breast cancer between the ages of 40 and 59, which indicates the prevalence of cancer in this younger demographic in particular (Ahmedin et al. 2004). Standard treatment models include a combination of surgery, chemotherapy, radiation and/or hormonal therapy. Surgical diagnosis is focused on the degree of breast cancer. Females are more likely to have two primary treatment alternatives for early-stage breast cancer: breast conserving surgery (BCS) and axillary staging, accompanied by adjuvant radiation (Fisher et al. 1995) or mastectomy and axillary staging, which is widely available.

The purpose of BCS is to remove pre-malignant and malignant lesions from the breast while reducing the alteration of the structure and volume of the breast. However, the cosmetic outcome obtained is less consistent, especially in smaller-breasted women. When all reconstructive alternatives are available, patients, including those qualifying for BCS, can consider progressing straight to mastectomy with immediate breast reconstruction (IBR). This method minimizes the risk of clinical recurrence and also the need for radiation treatment to the residual breast tissue thus instantly healing the breast.

Breast cancer therapy helps patients improve body image and quality of life are critical facets of patient-centered care, and IBR may play a significant part in this cycle (Ananian et al. 2004; Paulson et al. 1994; Gerber et al. 2009). Reconstructive pacing may be staggered (months or years after mastectomy) or instant. IBR has many benefits, including the potential to supply a woman with instant breast shape while preventing second anesthetic morbidity and has a positive effect on patient satisfaction and quality of life (Ananian et al. 2004).

As regards rapid autologous restoration, the protection of the native skin shell makes for fairly limited scarring and diminished adherence of the donor skin paddle around the chest wall (Kronowitz 2007). Patients intended for tissue expansion might have better results when the expander is implanted at the time of mastectomy, especially considering the possible need for postoperative radiation in locally advanced cancer cases (Kronowitz 2007; Koutcher et al. 2010; Alderman et al. 2000). Post-mastectomy radiation after IBR can also be used as an alternative to have sufficient local-regional power.

Despite these potential advantages, IBR is mostly limited to patients with low-risk diseases such as ductal carcinoma in situ (DCIS) and stage I disease (Alderman et al. 2000). For those with invasive breast cancer, prompt restoration is often prevented on the basis of the potential possibility that IBR can prolong adjuvant therapy, leading to adverse oncological outcomes (Kronowitz 2007).

The aim of this research was to systematically review the literature, comparing the incidence of local recurrence in mastectomy and IBR patients versus mastectomy alone in women with invasive breast cancer. The primary consequence was recurrence of breast cancer; the secondary finding was systemic recurrence.

Table 2.5. Comprehensive Analysis of ML Methods for Diagnosing Breast Cancer

Author & Year	Data Source	Machine Learning Technique	Accuracy
(Wang et al. 2020)	Wisconsin Diagnostic Breast Cancer (WDBC) dataset	Decision Table	92.92 %
		C4.5	93.45 %
		RIPPER	94.04 %
		OneR	88.57 %
		PART	93.95 %
		NB	89.46 %
		KNN	92.80 %
		SVM	97.66 %
		ANN	93.34 %
		RF	97.02 %
		Improved Random Forest-based Rule Extraction (IRFRE) [Proposed]	95.09 %
	Wisconsin Original Breast Cancer (WOBC) dataset	Decision Table	94.12 %
		C4.5	94.57 %
		RIPPER	95.39 %
		OneR	91.81 %
		PART	94.71 %
		NB	88.30 %
		KNN	96.58 %
		SVM	95.59 %

		ANN		96.30 %
		RF		97.01 %
		IRFRE [Proposed]		96.44 %
	Surveillance, Epidemiology and End Results (SEER) breast cancer dataset	Decision Table		81.71 %
		C4.5		75.5 %
		RIPPER		79.7 %
		OneR		79.65 %
		PART		77.28 %
		NB		60.1 %
		KNN		54.05 %
		SVM		79 %
		ANN		51.1 %
		RF		79.1 %
		IRFRE [Proposed]		80.45 %
(Alickovic & Subasi 2017)	Wisconsin Breast Cancer (Diagnostic) (WBC (DIAGNOSTIC)) dataset	Genetic algorithm-based feature selection	LR	98.45 %
			DT (C4.5)	94.02 %
			RF	95.43 %
			Bayes Net	95.34 %
			ANN (MLP)	98.45 %
			RBFN	94.38 %
			SVM	98.96 %
			Rotation Forest	99.48 %
(Goyal et al. 2020)	Breast cancer recurrence dataset from UCI Machine Learning Repository	Generalized Regression Neural Network (GRNN)		83.33 %
		Feed Forward Back Propagation Neural Network (FFBPN)		85.18 %
		Support Vector Machine (SVM)		77.77 %
		Decision Tree (DT)		70.83 %
		Naive Bayes (NB)		72.22 %
(Kumar et al. 2020)	Wisconsin Breast Cancer Diagnosis dataset from UCI repository	Naïve Bayes		73.21 %
		J48 Tree		82.81 %
		Ada Boost – M1		82.81 %
		Decision Table		83.09 %
		J – Rip		86.10 %

		Multiclass Classifier		94.27 %
		Logistics Regression		97.13 %
		Multilayer Perceptron		97.42 %
		Random Tree		99.14 %
		Random Forest		99.14 %
		Lazy K – Star		99.14 %
		Lazy IBK		99.14 %
(Islam et al. 2017)	Wisconsin Breast Cancer (WBC) dataset from UCI machine learning repository	SVM		98.57 %
		KNN		97.14 %
(Sakri et al. 2018)	Wisconsin Breast Cancer Prognostic Dataset	Naïve Bayes	Feature selection using Particle Swarm Optimization (PSO)	81.3 %
		Fast Decision Tree Learner (Reduced Error Pruning Tree Classifier - REPTree)		80.0 %
		KNN (IBK)		75.0 %
(Devi & Devi 2016)	Wisconsin Breast Cancer Dataset (WBC)	Farthest First clustering algorithm, Outlier Detection Algorithm (ODA), J48 classification algorithm		99.9 %
	Wisconsin Diagnosis Breast Cancer (WDBC)			99.6 %
(Chaurasia et al. 2018)	Breast Cancer Wisconsin dataset	Naïve Bayes		97.36 %
		RBF Network		96.77 %
		J48 Decision Tree		93.41 %
(Abdar et al. 2020)	Wisconsin Diagnostic Breast Cancer (WDBC) dataset	SV-BayesNet-2-MetaClassifier		97.72 %
		SV-BayesNet-3-MetaClassifier		98.07 %
		SV-Naïve Bayes-2-MetaClassifier		97.72 %
		SV-Naïve Bayes-3-MetaClassifier		98.07 %
(Asri et al. 2016)	Wisconsin Breast Cancer (original) dataset	C4.5		95.13 %
		SVM		97.13 %
		NB		95.99 %

		KNN		95.27 %
(Bhardwaj & Tiwari 2015)	Wisconsin Breast Cancer Dataset (WBC)	Genetically Optimized Neural Network (GONN)	50–50 training–testing data	97.73 %
			60–40 training–testing data	99.11 %
			70–30 training–testing data	99.21 %
			10-fold cross validation	99.26 %
(Wang et al. 2018)	Wisconsin Original Breast Cancer (WBC) Dataset	SVM-based Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE)		97.10 %
	Wisconsin Diagnostic Breast Cancer (WDBC) Dataset			97.68 %
	SEER Breast Cancer Dataset			76.42 %
(Boeri et al. 2020)	1021 patients who underwent surgery for breast cancer from April 2008 to December 2016 in SSD Breast Unit – ASST-Settelaghi Varese, Senology Research Center, Department of Medicine, University of Insubria, Varese, Italy	Loco-regional recurrence	Artificial Neural Network (ANN)	96.17 %
			Support Vector Machine (SVM)	96.86 %
		Systemic recurrence	ANN	95.29 %
			SVM	95.64 %
(Ibrahim & Shamsuddin 2018)	Breast Cancer Wisconsin Dataset	multilayer perceptron (MLP) neural network based on enhanced non-dominated sorting genetic algorithm (NSGA-II), Pareto optimal algorithm		97.01 %
(Kadam et al. 2019)	Breast Cancer Wisconsin (Diagnostic) medical data set (WDBC)	Proposed feature ensemble learning based on Stacked Sparse Autoencoders and Softmax Regression Model (FE-SSAE-SM model)		98.60 %

		SSAE-SM model		98.25 %
(Emami & Pakzad 2019)	Wisconsin Diagnostic Breast Cancer Dataset (WDBC)	Proposed Combined Affinity Propagation – Adaptive Modified Binary Firefly Algorithm (AP – AMBFA)		98.606 %
		AMBFA		98.21 %
		AP – BFA		98.54 %
		BFA		98.17 %
(Kanimozhi et al. 2019)	Wisconsin Original Breast Cancer Dataset	Intelligent Fuzzy Temporal Rule based Prediction Algorithm, C4.5 Decision Tree Algorithm		99.133 %
(Liu et al. 2019)	Wisconsin Original Breast Cancer (WBC) Dataset	Information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection + BP neural network		96.3 %
		IGSAGAW + 3-NN		95.6 %
		IGSAGAW +cost sensitive support vector machine (CSSVM)		95.8 %
	Wisconsin Diagnostic Breast Cancer (WDBC)Dataset	IGSAGAW + BP		97.5 %
		IGSAGAW + 3-NN		95.4 %
		IGSAGAW +CSSVM		95.7 %
(Mohebian et al. 2017)	Breast Cancer Dataset from Isfahan Sayed-o-Shohada cancer research center	HPBCR (hybrid predictor of breast cancer recurrence) with Particle Swarm Optimization (PSO)& Bagged Decision Tree (BDT)		90 %
		HPBCR with BDT excluding PSO		78 %
		HPBCR with SVM & PSO		90 %
		HPBCR with MLP & PSO		88 %
(RamaDevi et al. 2018)	Breast cancer (BC)Dataset	PCA + SMOTE (%) = 75	Ensemble model of KNN + RF	97.99 %

	Wisconsin breast cancer (WBC) Dataset	PCA + SMOTE (%) = 90	Ensemble model ofKNN + RF	100 %
	Wisconsin diagnostic breast cancer (WDBC) Dataset	PCA + SMOTE (%) = 68	Ensemble model ofLR + KNN	98.32 %
	Wisconsin prognostic breast cancer (WPBC)Dataset	PCA + SMOTE (%) = 100	Ensemble model ofKNN + RF	100 %
(Nilashi et al. 2017)	Wisconsin Diagnostic Breast Cancer (WDBC) Dataset	EM-PCA-CART-Fuzzy Rule Based Method		93.2 %
	Mammographic Mass Dataset			94.1 %
(Polat & Sentürk 2018)	Breast Cancer Coimbra Dataset	MAD (median absolute deviation) normalization + K-means clustering based feature weighting (KMCBFW) + Adaptive Boosting (AdaBoostM1) classifier		91.37 %
(Sharma et al. 2017)	Wisconsin diagnostic breast cancer (WDBC) Dataset	Logistic Regression		96.89 %
		k-Nearest Neighbor		93.06 %
		Support Vector Machine		89.6 %
	Wisconsin prognostic breast cancer (WPBC) Dataset	Logistic Regression		88.6 %
		k-Nearest Neighbor		82.56 %
		Support Vector Machine		89.73 %

Sharable Data Are Key

Data can come from several sources. The research papers studied in this chapter are not unique to one particular dataset, or even one form of data. Researchers have applied ML effectively to clinical notes (Rumshisky et al. 2016), physiological waveforms (Shoeb and Guttag 2010; Saria et al. 2010), standardized Electronic Health Record (EHR) data (Ghassemi et al. 2015), radiological images (Ahmed et al. 2016) and even unstructured journal data. The degree of incompatibility, inaccuracy, and error, specifically in clinical reports, is commonly known to healthcare providers. Mostly, the vast majority of such projects concentrate on "data wrangling" i.e., data retrieval and pre-processing.

Shared data sets accomplish an essential function by making it possible to compare standard ML procedures with particular health conditions. It is difficult to compare methods in a meaningful way without a shared dataset.

When selecting a target (i.e., a result of interest), one must have access to reliable data on that target. For example, to create this model, if the goal is to forecast the progress of CVD, it must be understood which patient developed CVD. At times, total consistency cannot be obtained (like all laboratory experiments are not 100% accurate). However, the possibility of any vulnerability in the data may be addressed by ML techniques. Furthermore, we must note that the consequence that the model learns to expect during the training is the outcome. For instance, we would like to predict CVD risk. However, because not all patients have been tested, we may simply estimate the probability of a positive laboratory outcome for CVD. This distinction is subtle, but it is essential. In particular, if the hospital changes its test procedure, the predictive output of an existing model can change (Wiens and Shenoy, 2018).

The most appropriate methods found in the literature for the treatment of missing data are explained below (García-Laencina et al. 2015):

Imputation-based methods: The missing data is estimated and filled with plausible values using the available full data. The classification is then designed with the data set imputed. Two separate and consecutive phases, imputation and classification, could therefore be taken into account.

Avoid explicit-Imputation: Resolution methods can deal with unknown values in these types of approaches, i.e., they can handle missing values during the design of

classification strategy. The classification is then carried out without imputation of previously missing data.

Usually, choosing appropriate methods for a decision-making process is a challenging task, which leads to the best results for classification. This could become more challenging if the final decision is extremely crucial, for example in the field of healthcare. Two possible strategies can be followed to select the best decision in this field. Firstly, the classification technique generally reaches higher accuracy than other techniques regardless of the problem of decision making. The second is to obtain the classifier that exceeds the accuracy of the remaining classifiers concerning problem specifications. As no classification method in machine learning is better than other techniques, the only practical approach is the second option (Baati et al. 2016). For example, Deep Learning (DL), which has demonstrated promising abilities in automatically detecting complex, patterns with high volumes and varieties of data. As such, DL can complement physical models in the modeling of complex disease detection processes to better understand changes in process output (Wang et al. 2019). In Table 2.6., various datasets used in the literature are described.

Table 2.6. Description of Datasets used in This Study

Dataset	Source	Total Number of Records	Instances of Missing Values	Records with Disease	Records without Disease	Number of Attributes	Number of Used Attributes
UCI Machine Learning Dataset Repository – Heart Disease	UCIMLR	303	3	–	–	76	14
Cleveland Heart Disease	UCI-A	303	6 missing values	139	164	76	13 or, 14
Hungarian Heart Disease	UCI-B	294	782 missing values	106	188	76	13 or, 14
Statlog Heart Disease	UCI-C	270	0	120	150	76	13
SPECTF Heart Disease	UCI-D	267	–	212	55	45	45
Long Beach Heart Disease	UCI-E	200	3	149	51	76	13 or, 14

Switzerland Heart Disease	UCI-F	123	273 missing values	115	8	76	13 or, 14
Korea National Health and Nutrition Examination Survey (KNHANES)	KNHANES	25990	–	12915	13075	14	14
Kaggle Framingham Heart	Kaggle FHS	4238	582 (rows)	557	3099	16	16
Framingham Heart Study	FHS	4240 – 4434	–	644 (in 4240)	3596 (in 4240)	16 – 22	–
NHANES Physical Activity and CVD Fitness Data	NHANES	–	–	–	–	–	–
Coronary Heart Disease Data from Iranian Hospitals	(Sabahi 2018)	152	–	–	–	–	–
Z-Alizadeh Sani	(Arabasadi et al. 2017)	303	–	216	87	54	54
Patients' medical data from Enam Medical Diagnosis Centre, Savar, Dhaka, Bangladesh	(Sultana et al. 2016)	100	–	–	–	6	6
Apollo Hospitals CKD Dataset	CKD-A	400	242 incomplete instances	250	150	25	25
CKD Dataset from University Hospital located at the Federal University of Alagoas (UFAL), Brazil	(Sobrinho et al. 2020)	60 (real world), 54 (augmented)	–	44 (in 60)	16 (in 60)	8	8

Ambulatory Electronic Medical Record (EMR) of Patients from Tawam Hospital, Al Ain, UAE	(Al-Shamsi et al. 2018)	544	number of used records = 470	416 (advanced stage)	54 (early stage)	–	–
CKD Dataset from National Kidney Foundation, Bangladesh	(Saha et al. 2019)	13000	–	8125	4875	26	26
[Source Unknown]	(Ripon 2019)	2800	–	1750	1050	24	24
Saudi CKD Dataset retrieved from King Fahd University Hospital (KFUH), Khobar	(Alassaf et al. 2018)	244	–	118	126	57	57
Pima Indians Diabetes Database	PID	768	0	268	500	8	8
Localized Diabetes Dataset	(Choubey et al. 2020)	1058	0	753	305	12	12
Diabetes [2009–2012] Dataset from National Health and Nutrition Examination Survey (NHANES)	NHANES-A	6561 (excluding missing values)	–	657	5904	14	14
Primary Clinical Dataset from a Diagnostic Lab in Kashmir Valley	(Shuja et al. 2020)	734	–	430	304	11	11
Pima Indians Diabetes Database from National Institute of	(Yuvaraj and SriPreethaa 2019)	75664	–	–	–	13	13

Diabetes and Digestive Diseases							
2014 BRFSS (Behavioral Risk Factor Surveillance System, 2014) Data from CDC	BRFSS	138146 (out of 464644)	–	20467 (in 138146)	390827 (in 464644)	279	27
Annual Physical Examination Reports from Shengjing Hospital of China Medical University, Liaoning Province	(Pei et al. 2019)	4205 (out of 8452)	3956 (out of 8452)	709	3496	9	10
Physical Examination Data (Diabetic & Non-Diabetic Patients) from Nanjing Drum Tower Hospital, China	(Xiong et al. 2019)	11845	–	3845	8000	450	11
NHANES 2013-14 Diabetes Dataset	NHANES-B	10172	–	–	–	54	24
Dataset from Sawanpracharak Regional Hospital, Thailand	(Nai-arun and Moungmai 2015)	30122	–	10977 (in risk)	19145	11	11

Discussion & Analysis of ML Techniques

For the diagnosis of CVD, CKD, and Diabetes, numerous machine-learning algorithms are extensively used. Assessment of existing literature reveals that LR, NB, SVM, KNN, DT, and RF are widely used state of the art algorithms for the detection of diseases. There are a variety of solutions to various ML methods in smart healthcare. However, there is indeed a "No Free Lunch" theorem in ML that says no single algorithm delivers the best result for every application (Wolpert 2002). As a result, as people develop a smart healthcare application algorithm, the research is time-consuming because one can go through several algorithms and pick the one with the best output. The most crucial know-how is to test out the best algorithms for the problem. For example, clustering algorithms do not seem to unravel classification problems. In recent years, Deep learning has earned attention, owing to its dominance in solving complex problems. Nevertheless, it has been recommended that deep learning is not necessarily the right or the most appropriate solution for every problem. It depends on the intricacy of the problem, the quantity of data available, the computing capacity, and the training time (LeCun et al. 2015).

Existing literature suggests RF and DT algorithms are more efficient than other algorithms. K Nearest Neighbor is also very much useful for prediction purposes. It also displays the maximum accuracy, but it takes more time in comparison to other algorithms. Tree algorithms are also used but have not been widely accepted due to their complexity. They also show enhanced accuracy when responding correctly to data set attributes.

Among ensemble learning techniques, RF Classifier is very popular and efficient in DM and ML for high-dimensional classification and skewed problems (Azar et al. 2014). The high variation is the weakness of tree classifiers. Besides, it is not unusual for a minor alteration in the training data set to create a vastly separate tree. The explanation for this is the hierarchical existence of the tree classifiers. In a tree, an error that arises in a node near the root spreads to the leaves. A decision on forest methodology has been invented to make tree classification more stable. A decision forest is a cluster of decision trees. It may use as a standard classifier that includes multiple classification methods or a single method with separate operating parameters. Consider the $S = ((X1, Y1), ..., (Xk, Yk))$ made up of k-vectors, $X \in P$ where P contains a set of numerical or symbolic observations, and $Y \in Q$ where Q is a set of class labels. In the case of classification problems, the

classifier is mapping P → Q. Each tree in the forest is classified as a new input vector. Every tree yields the product of a certain assignment. The theory of random forests is to create binary sub-trees using training bootstrap samples from the learning sample L and randomly select a teaching node from a subset of X.

The decision forest selects the classification that has the most votes of all the trees in the forest. Random forest technique involves Breiman's "bagging" concept and Ho's (Ho 1995) "random selection features." Bagging, which stands for "bootstrap aggregation," is a method of ensemble learning developed by Breiman (Breiman 1996) to increase the accuracy of a poor classifier by generating a collection of classifiers. Random forest uses ensemble method that incorporates the predictions of several individual tree models (base classifiers) to produce a prediction that appears to be more reliable than all of the individual classifiers' predictions. The results showed that the RF Classifier had the highest accuracy in CVD, CKD, and Diabetes predictions and the values were 99.6%, 100%, and 100% respectively. This work addresses 184 articles on the use of techniques in clinical decision-making.

A. Discussion on Reviewed Papers

In this review article, the authors noticed that the accuracy and efficiency of various data mining techniques vary based on the nature of features present in the datasets and the size of different training and testing partitions of datasets. Generally, healthcare datasets are highly imbalanced in nature. Classification of these datasets results in erroneous prediction and inaccurate accuracies. Another very common characteristic of healthcare datasets is a large number of missing data values for multiple features. The sample size of the data is also seen as a different aspect, as the data available are typically limited in size. No single data mining techniques can resolve all the aforementioned issues independently. We have noted some more observations related to all the reviewed papers about different MLTs used in this study. These are illustrated below:

1) Decision Tree

After comparing various research works authors can conclude that DT cannot be used to solve prognostic decision problems for imbalanced datasets because DT splits observations into branches recursively to construct the tree.

2) K-Nearest Neighbor

This is a very expedient algorithm as it permanently stores the information present in the training dataset. But because of its time-consuming nature, this algorithm is only suitable for large data sets. At the time of classifying a new dataset, this algorithm needs a longer classification time to process individual data in the training dataset. In this study, the authors noticed that the accuracy of the classification is what the researchers would like to achieve instead of the time of classification since the accuracy of the classification is more important in the medical diagnosis.

3) Logistic Regression

Due to the significant decrease in the size of input datasets the outcome of LR analysis on different datasets by various researchers was not very significant. An LR technique is most suitable for large datasets. If the datasets were large as the boundaries of precision were larger, the results would be more significant.

4) Naïve Bayes

The NB classifier is very much efficient to handle missing values present in datasets and at the same time, it is computationally very competent also. Based on these advantages most of the researchers recorded a high value of accuracy for the prediction of diseases. NB also allows the researchers to the extraction of more features without overfitting the datasets. NB would be a very efficient approach for the datasets with a large number of missing values.

5) Support Vector Machine

There are two different factors to control the generalization ability of the SVM method: (1) the error in the training and (2) the capacity of the learning machine. By changing the features in the classifications, the error rate can be controlled. It is clear from the results obtained from the studies that the SVM has shown greater efficiency since it maps the features into higher dimensional space.

6) Random Forest

The RF classifier is one of the most successfully implemented ensemble learning techniques which have proved very popular and powerful for high-dimensional classification and skewed problems in pattern recognition and ML. It offers the benefit of

computing efficiency and improves the accuracy of predictions without considerably increasing calculation costs. Based on these characteristics most of the researchers recorded the highest value of accuracy for the prediction of diseases.

B. Accuracy and Consideration of Bias-Variance Tradeoff

Accuracy indicates the ratio of the estimated value to the real or actual value (Tazin et al. 2016). The basic formula for accuracy is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FN}) + (\text{FP} + \text{TN})}$$

Where Total Positive = True Positive (TP) + False Negative (FN), Total Negative = False Positive (FP) + True Negative (TN), TP = Number of true samples in the dataset classifies as true, TN = Number of false samples in the dataset classified as false, FN = Number of true samples in the dataset classified as false, FP = Number of false samples in the dataset classified as true.

ML techniques with the highest accuracy are chosen for prediction in the classification system.

Under-fitting occurs in supervised ML techniques when an ML model cannot capture the underlying pattern of the dataset. These ML techniques are typically highly biased and have low variance. It happens when a researcher(s) has very few records in the dataset to build a precise model or when they are trying to build a linear model with nonlinear data. Such models are also very easy to capture complex patterns in data such as LR and Linear regression. Over-fitting occurs when ML algorithms capture the noise in the datasets along with the underlying pattern. It happens if someone highly trains the ML algorithm over noisy datasets. These ML techniques are of low bias and high variance. These models are very complex like DTs that tend to over-fit.

If the ML model is very basic and has too few features, it may be very biased and of little variance (Picon et al. 2019). If the ML model has more features, on the other hand, it will be strongly variant and low in bias. So, without over-fitting and under-fitting data, we need to find the correct or good balance. To handle the situation a tradeoff in complexity has to be made by which the balance between bias and variance can be maintained. An

algorithm cannot simultaneously be more complex and less complex (Défossez and Bach 2015). To develop an effective ML model, the researcher must find a good balance between bias and variance, so that the total error is minimized.

C. Findings with Timeline

Moreover, we have considered articles published between 2010 and 2020. The purpose of the study is to provide a detailed analysis of the use of ML and DM techniques in the field of clinical decision-making and the application of the most commonly used state of the art classifiers. While this study cannot be considered rigorous and comprehensive, it still provides a general background and description of work in this field and offers valuable guidelines for researchers in this field. The findings reported in this chapter have several significant implications:

The total number of research papers considered for CVD is 34. Out of these 34 articles, 4 papers are used for classification, 4 are used for diagnosis, 23 papers are related to prediction, 2 of them address both prediction and classification, and 1 is related to prediction and diagnosis. For all 34 research papers in 20 journals, only state-of-the-art classifiers are used, 16 journals use NN and the remaining 18 papers deal with hybrid/ensemble techniques.

For CKD, the aggregate of research papers listed is 39. Out of these 39 papers, 2 papers are used for classification, 12 are used for diagnosis, 22 for prediction, 2 for prediction and classification, and 1 for prediction and diagnosis. Of those 39 research papers in 27 journal articles, only state-of-the-art classifiers are used, 18 articles use NN and the remaining 13 papers deal with hybrid/ensemble techniques.

In the case of Diabetes, the total number of research papers considered is 45. Out of these 45 articles 5 papers are used for classification, 11 are used for diagnosis, 24 papers are related to prediction, 1 of them discusses the prediction, as well as classification, and 4 is related to prediction and diagnosis. Among those 45 research papers in 26 papers, only the state-of-the-art classifiers are used, 18 papers use NN, and the rest 21 papers deal with Hybrid/Ensemble techniques. Year-wise distribution of classification techniques and working principles are illustrated in Figures 2.2 and 2.3 respectively. The timeline of ML techniques based on maximum classification accuracies for CVD, CKD, and Diabetes are

tabulated in Table 2.9., 2.10., and 2.11. respectively.

		SVM	RF	KNN	NB	DT	LR	NN
Total Literature Tabulated (118)	Heart Disease (34)	18 (53%)	11 (32%)	14 (41%)	16 (47%)	16 (47%)	8 (23%)	16 (47%)
	Chronic Kidney Disease (39)	21 (54%)	15 (38%)	14 (36%)	17 (43%)	16 (41%)	8 (20%)	13 (33%)
	Diabetes (45)	17 (38%)	11 (24%)	11 (24%)	20 (44%)	18 (40%)	12 (27%)	12 (27%)

Figure 2.2. Distribution of Methods used to Classify Different Diseases in Recent Years

		Classification	Diagnosis	Prediction	Prediction & Classification	Prediction & Diagnosis
Total Literature Tabulated (118)	Heart Disease (34)	4 (12%)	4 (12%)	23 (68%)	2 (6%)	1 (3%)
	Chronic Kidney Disease (39)	2 (5%)	12 (31%)	22 (56%)	2 (5%)	1 (2%)
	Diabetes (45)	5 (11%)	11 (24%)	24 (53%)	1 (2%)	4 (9%)

Figure 2.3. Distribution of Literature Based on Working Principle in Recent Years

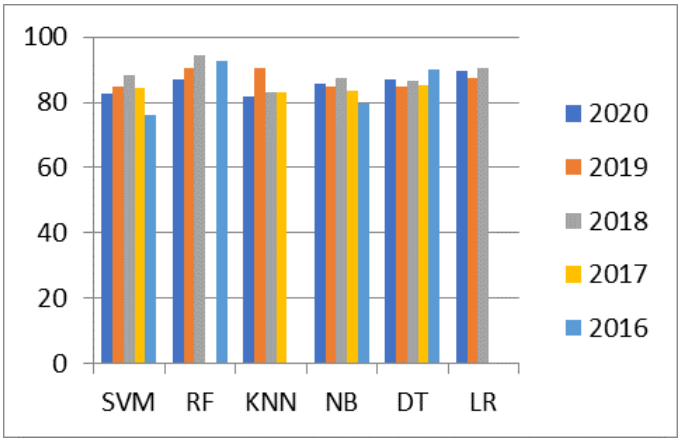


Figure 2.4. Accuracies of Various ML Methods for Diagnosing Heart Disease

Table 2.7. Analysis of Various ML Methods for Diagnosing Heart Disease

ML Techniques	Average Accuracy				
	2020	**2019**	**2018**	**2017**	**2016**
SVM	82.49	84.7	88.39	84.15	76.07
RF	86.86	90.37	94.3		92.44
KNN	81.94	90.45	83.24	83.16	
NB	85.67	84.68	87.58	83.49	79.63
DT	86.91	84.62	86.43	85.07	89.96
LR	89.59	87.57	90.29		
NN	90.28	89.89	89.46	86.34	

Table 2.8. Analysis of Various ML Methods for Diagnosing Chronic Kidney Disease

ML Techniques	Average Accuracy				
	2020	**2019**	**2018**	**2017**	**2016**
SVM	92.51	96.24	97.79	98.11	98.40
RF	91.77	98.7	92.54	99.42	
KNN	87.16	84.59	96.79	97.69	97.80
NB	88.61	98.38	97.17	98.12	97.50
DT	92.19	97.03	96.02	99.05	96.90
LR	90.2	98.02	98.1	95	96.55
NN	90.4	89.97	99.07	98.44	

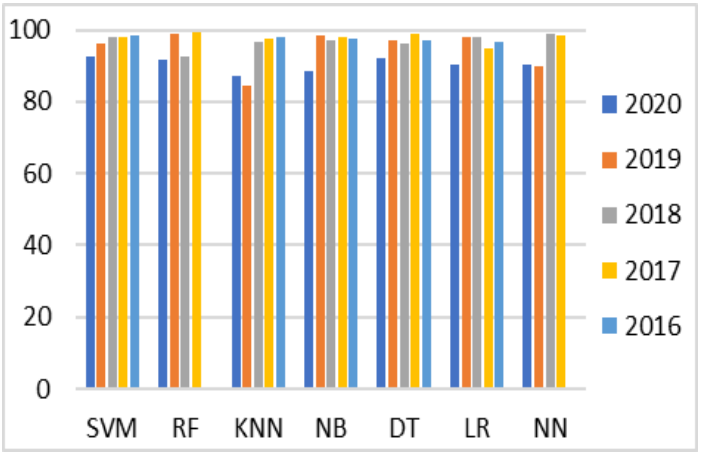


Figure 2.5. Accuracies of Various ML Methods for Diagnosing Chronic Kidney Disease

Table 2.9. Analysis of Various ML Methods for Diagnosing Diabetes

ML Techniques	Average Accuracy				
	2020	2019	2018	2017	2016
KNN	80.15	76.17	78.83		
DT	84.65	78.86	75.91		
SVM	89.88	79.70	78.32		
LR	87.48	78.85	77.74		
NN	93.65	80.4	80.80		
NB	85.64	80.2	77.18	77.74	
RF	93.84	81.51			

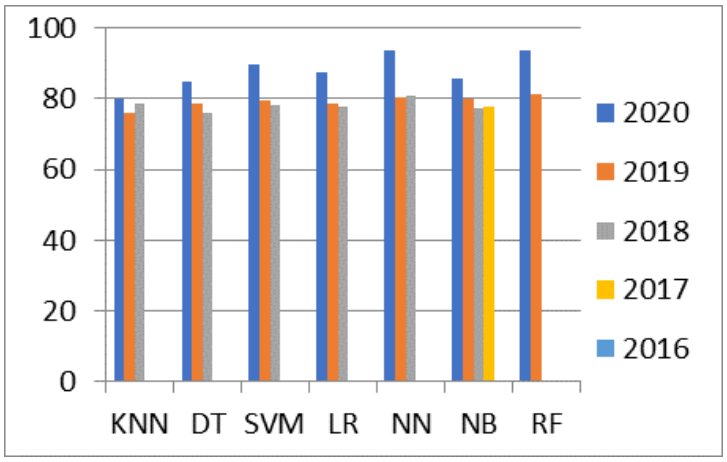


Figure 2.6. Accuracies of Various ML Methods for Diagnosing Diabetes

Table 2.10. Timeline of ML Techniques Based on Maximum Classification Accuracies for CVD

Year	Techniques Used	Max Accuracy		
		Reference	Techniques Used	Accuracy (%)
2020	KNN, CART, SVM, NB, NN, RF, DT, LR, Extra Trees, ACO, FA, PSO, ABC, LM, LDA, SAE	(Ayon et al. 2020)	NN	98.15
2019	FCBF, PSO, ACO, KNN, SVM, RF, NB, MLP, NN, GA, DT, LR, Vote, PCA, LDA, RFSA, FSSA, FFSA, MFSFSA, RBF, SGB	(Khourdifi and Bahaj 2019)	KNN	99.65

2018	GA, MDMS-PSO, NB, Bagging, DT, LR, KNN, RF, SVM, NN, RIPPER, BFAHP	(Rajliwall et al. 2018)	NN	98.8
2017	GA, NN, RFRS, DT, DE, AHP, NB, KNN, MLP, RBF, SCRL, SVM	(Uyar and Ilhan 2017)	GA + RFNN	96.63
2016	DT, NB, RF, SVM, Discriminant, KStar, SMO, Bayes Net, MLP	(Zriqat et al. 2016)	DT	99.01
2015	IT2FLS, BPSORS-AR, CFARS-AR	(Long et al. 2015)	IT2FLS + CFARS-AR	88.3

Table 2.11. Timeline of ML Techniques Based on Maximum Classification Accuracies for CKD

Year	**Techniques Used**	**Max Accuracy**		
		Reference	**Techniques Used**	**Accuracy (%)**
2020	Fruit Fly Optimization, SVM, KNN, DT, LR, NN, NB, RF, MLP	(Rubini and Perumal 2020)	Fruit Fly Optimization + MKSVM	98.51
2019	DFS, ACO, LOG, RF, KNN, SVM, NB, LDA, AdaBoost, LogitBoost, XGBoost, Extra Trees, LR, DT, NN, SMO, IB1, VFI, ZeroR, MLP, Decision Stump, REPTree, CART, RFE, ETC, US	(Rabby et al. 2019)	GNB	100
		(Rabby et al. 2019)	DT	100
		(Pasadana et al. 2019)	RF	100
		(Ogunleye and Qing-Guo 2019)	RFE + ETC + US + XGBoost	100
2018	XGBoost, SVM, C&R Tree, CHAID, Quest, DT, KNN, NN, NB, RBF, RPART, LR, MLP, RF, MLP-FFN, NN-GA	(Hore et al. 2018)	NN-GA	100
		(Alaoui et al. 2018)	XGBoost Linear	100
2017	SVM, NN, NB, DT, KNN, Back Propagation, MLP-FFN, NN-GA, NN-CS, LR, SMO,	(Sisodia and Verma 2017)	RF	100

	RF, Bagging, AdaBoost, CFS, REPTree, BFTree	(Subasi et al. 2017)	RF	100
		(Basar and Akan 2017)	BFTree + RF + Information Gain Attribute Evaluator Feature Selection	100
2016	NB, RA, SVM, DT, KNN, LR, FuRES, FOAM, PLS-DA	(Chen et al. 2016)	FuRES	99.6
2015	WrapperSusetEval Attribute Evaluator, NB, SMO, IBK, Best First Search	(Chetty et al. 2015)	WrapperSusetEval Attribute Evaluator + IBK + Best First Search	100

Table 2.12. Timeline of ML Techniques Based on Maximum Classification Accuracies for Diabetes

Year	**Techniques Used**	**Max Accuracy**		
		Reference	**Techniques Used**	**Accuracy (%)**
2020	NSGA-II, LR, KNN, DT, NB, PCA, PSO, AdaBoost, RF, Bagging, SVM, MLP, LR, SMOTE, SMO	(Devi et al. 2020)	FF Clustering + SMO	99.4
2019	DT, NB, RF, ANT-FDCSM, SVM, KNN, DT, NB, LR, MLP, BPNN-CSA, LVQAC, MODLNN, CTCPNN, NNGE, NN, AdaBoost, Bayes Net, REPTree, LMT, GTB, RF-WFS, XGBoost, RST, RBFNN-HS, LogitBoost, RobustBoost, Bagging, GA, GRNN O., R. GRNN O., Back Propagation, LDA, ACO, GWO	(Gbengaa et al. 2019)	NNGE	100
2018	SVM, RBF, KNN, NN, MDR, NB, DT, K-means, LR, RF, CART, MMS	(Husain and Khan 2018)	KNN	96.1
2017	PCA, SOM, NN, K-Means, GA, SVM, DT, SW FFANN, NB, GMM, ELM, LR, AutoMLP, LDA, QDA, GPC	(Bhatia and Syal 2017)	K*-Means + GA + SVM	97.959

2016	DT, KNN, PCA	(Panwar et al. 2016)	KNN + PCA	100
2015	DT, SVM, NB, AdaBoost, KNN, RF, NN, LR, Bagging, Boosting	(Kandhasamy and Balamurali 2015)	KNN	100
			RF	100

Chapter 3

Machine Learning Algorithms and Performance Metrics

Discriminant Analysis

Discriminant analysis uses continuous variable measurements on diverse associations of items to focus aspects that differentiate the groups and to use these appraisals to classify new items. Discriminant function analysis is a parametric technique to determine the weightings of quantitative variables or seer variables best discriminate between two groups of seer variables or more than two groups of seer variables (Cramer 2003). The analysis generates a discriminant function which is nothing but a linear amalgamation of the weights and significant variables.

The primary reason to use multiple linear regressions is to define a function where one dependent variable is linear combination of independent variables. In discriminant analysis a single qualitative dependent variable is a linear weighted sum of independent variable(s). The dependent variable comprises two classes or combinations in most of the cases, like, existence or non-existence of cardio vascular disease, high blood cholesterol versus normal cholesterol, recurrence and non-recurrence of malignancy in breast cancer patients etc. This combination of independent variables is such that it will discriminate best among the groups. This linear combination defines a discriminant function. In this discriminant function the association of the variables is generated through correction of weights, referred as discriminant coefficients, assigned to the independent variables.

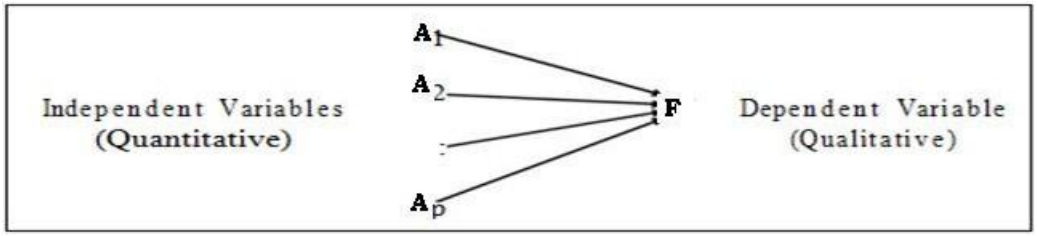


Figure 3.1. The Model

The discriminant equation:

$$F = \quad 0 + 1A1 + 2A2 + \ldots + \quad pAp + \varepsilon \text{ or } F = \sum Ai + \varepsilon \quad \ldots\ldots \quad (3.1)$$

Where, F is a dependent latent variable; Ai represents independent variables, A1, A2, … Ap; ε is the error term; and 0, 1, 2, …, p are the discriminant coefficients.

Assumptions are the variables A1, A2, …, Ap are not correlated. Sizes of the mutually discriminating Groups are not exclusively different. The maximum number of independent variables will be two less than the total sample size. The independent variables within each group of the dependent variable have similar variance-covariance structures. By nature, the errors (residuals) are distributed and organized arbitrarily. The distinct independent variables follow multivariate normal distribution. There are several purposes for MDA (Multiple Discriminant Analysis). Scrutinizing the differences between groups and identify the different groups, discarding variables which are less related to the group. Hence MDA helps achieve parsimony without affecting the accuracy of predictability. Creating groups by classifying different cases. The accuracy of the discriminant function is tested by observing whether test cases are classified as predicted. It works on already defined datasets for which the groups are defined earlier.

Determine secondary relationships between a group of discriminating variables and one classification variable; contemplating to derive a correlation between dependent and independent variables.

Take out prevalent, canonical functions within the group of adjunct variables from a number of illustrations, such that discrepancy among the group is increased and discrepancy within the group is decreased towards the gradient.

An adequate number of primary variables can be condensed together to create a small set canonical function by which dimensionality of a multivariate data will be reduced. Depending upon a collection of discriminating characteristics the absolute inequalities between the already defined groups of sampling entities has been described.

Logistic Regression (LR)

The LR classification defines the logistic model parameters (a form of binary regression) and is a specific type of linear regression model, though the dichotomous response factors violate the presumption of normality in generalized regression algorithms. The LR model ensures that the linear representation of the available explanatory variables' observed values is a proper function of the fitted probability of the case. LR's main potential is to generate a simple form of probabilistic classification (Ann et al. 2013).

The drawbacks are that LR cannot adequately address the problems of explanatory variables with non-linear and interactive results. LR is a type of regression used for estimating a binary dependent variable. It uses the maximum likelihood ratio in generating the LR equation to determine the statistical significance of the variables. Equation 3.2 describes the LR model.

$$prob\left(K=1\right)=\frac{1}{1+e^{-\left(\beta_0+\beta_1 a_1+\ldots\ldots+\beta_m a_m\right)}} \quad \ldots\ldots\ldots\ldots (3.2)$$

Where, prob(K=1) is the probability of existence of the disease and β_0, β_1, $\ldots\ldots\beta_m$ are coefficients of regression.

Within the logistic regression method, there is a linear model secret. The normal logarithm of the prob(K=1) ratio gives a linear model in a_i to (1 - prob(K=1)) as shown in equation 3.3 and expanded in linear form as shown in equation 3.4.

$$f\left(a\right)=ln\left(\frac{prob\left(k=1\right)}{1-prob\left(k=1\right)}\right)\ldots\ldots\ldots\ldots\ldots\ldots (3.3)$$

$$=\beta_0+\beta_1 a_1+\ldots\ldots..+\beta_m a_m \quad \ldots\ldots\ldots (3.4)$$

LR is useful in situations where one can predict the presence or absence of a dependent feature or outcome based on independent feature set values. It is similar to a linear regression variant but is suitable for models with a dichotomous dependent variable.

Naïve Bayes (NB)

NB classifier is Bayes theorem-dependent and deals with basic probabilistic classification with strong independent assumptions. The presence or absence of any particular variable depends, according to this classifier, on the presence or absence of other variables in the dataset (Patil & Sherekar 2013; Dimitoglou et al. 2012). NB classifier deals mainly with conditional probabilities. Bayes' theorem used a formula to calculate likelihood by counting in historical data the number of values and value combinations. Bayes' theorem calculates the probability that an occurrence will occur, given the likelihood that another event will occur already. If M is the dependent event based on the event N, Bayes' theorem can be stated as follows.

The algorithm counts the number of cases in which M and N coincide, dividing them by the number of cases in which M occurs alone. The benefit of the Naive Bayes classifier is that it needs a small amount of training data to estimate the parameters (means and variances of variables) required for classification. Since independent variables are known, only the variances of the variables shall be determined for each class and not the sum. It proposed for classification of multiple class issues.

It is a pretty fast ML algorithm and generally used in text mining environments like filtering spam, new article classification, and sentimental analysis (Dey, Hossain & Rahman 2018; Diab & El Hindi 2017). It can be defined mathematically as

$$P(M \mid N) = \frac{P(N \mid M)P(M)}{P(N)} \quad \text{..... (3.5)}$$

Where, $P(M \mid N)$: Represents probability of incident of event M given the event N is true.;

$P(N \mid M)$: Represents probability of the occurrence of event N given the event M is true.;

$P(M), P(N)$: Re *presentsp*robabilities of the occurrence of event M and N respectively

Support Vector Machine (SVM)

SVMs are one type of efficient ML technique with a high generalization capacity in practice. They are a group of margin classification models suggested by Vapnik and his group at AT&T Bell Laboratories in the 1990s (Cortes & Vapnik 1995). In contrast to the methods of statistical learning based on empirical risk minimization, the objective of SVM is to minimize structural risk, which demonstrates a strong ability to avoid over-fitting (Ayat, Cheriet & Suen 2005). In the SVM model, a decision hyperplane is used for a separation gap that divides the maximum margin between two classes. Compared to traditional ML approaches, SVMs have been applied to many fields for their widespread generalization abilities. In particular, as a data-driven prediction technique, in recent years, SVM models have drawn the most attention to the diagnosis of diseases, such as diagnosis of cerebral palsy gait, detection of gastric lymph nodes, and diagnosis of prostate cancer (Kamruzzaman & Begg 2006; Son et al. 2010; Ishikawa et al. 2014; Shah et al. 2012).

Decision Trees

The decision tree is a function that constructs tree-like classification structure (Mendonça et al. 2007). It segregates essential facts into classes. It also anticipates the values of a target or dependent variable constituted with the help of predictor or independent variables. These functions also produce authorization tools for preliminary and affirmatory study for classification. In our present paper, we are considering the results of prognosis for breast cancer to predict the chances of recurrences in data mining by using decision tree approach.

We are using two most important growing methods in decision tree algorithm that are Chi-squared Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) analyses. These two algorithms build two different types of tree structures (Lakshmanan et al. 2015). Collections of “if-then-else” rules are the main constructor of decision tree algorithm that shows the instructions in exhaustive form for several cases. In thismethod, conclusive tree will be the result that is produced from categorical input data. For categorical huge datasets, CHAID is treated as the best method among all other decision

tree algorithms (Hastie et al. 2009). Though these two tree structures have a number of differences between them, they can be used for same purposes also.

Especially for analyzing very large datasets, CHAID provides a much better tree structure than compared to CART (Cao et al 2010). CHAID (Tu et al. 2009) was designed to analyze the definitive and discontinuous target datasets, and it can break the data by using multi-way split technique; it is actually splitting of present node into more than two nodes automatically. CHAID is actually a pattern of decisive test by which relationships between dependent and independent variables can be described. The routine statistical tools like regressions cannot be used with breast cancer datasets (Howlader et al. 2015). CHAID is considered to be the perfect technique to determine the relationships between dependent and independent variables because it is more convenient to investigate the categorical data (Mendonça et al. 2007; Liu et al. 2009).

Random Forest (RF)

RF is a well-known supervised classification method used in various classification fields. It is an ensemble learning technique (Loh 2011) that operates on the concept of using a collection of weak learners to prepare a strong learner. RF uses the Classification and Regression Tree (CART) techniques (Breiman 2001) to create a mixture of multiple decision trees based on the bootstrap aggregation (bagging) technique (Liaw & Wiener 2002). The CART methodology correctly classified the dependent and independent variables and creates a relationship between them. In RF, each tree randomly chooses a subset of the dataset to build an independent decision tree. RF splits the selected random subset from the root node to the child node repeatedly until each tree reaches the leaf node without pruning. Each tree independently classifies the features and the target variable and votes for the final tree class. Depending on the majority of the votes cast, RF decides the final overall classification.

Extra Trees (ET)

A variant of a random forest is an "extra trees" classifier (ETC), also known as an "Extremely randomized trees" classifier. The entire sample is used at each step in an ET, unlike a RF, and the decision boundary is randomly chosen rather than the best. In real-world cases, performance is comparable to an ordinary, RF, sometimes a bit better (Landwehr et al. 2005; Sharaff & Gupta 2019). In particular, it is an ensemble of decision trees and is associated with other algorithms for decision trees, such as bootstrap aggregation (bagging) and RF. By generating a large number of un-pruned decision trees from the training dataset, the ET algorithm operates. In the case of regression, predictions are made by averaging the prediction of the decision trees or by utilising the majority vote in the case of classification.

Gradient Boosting (GDB)

One of the best-supervised machine learning algorithms for regression and classification problems is GDB. The GDB algorithm is likely to make decisions in the form of an ensemble of weak prediction models. It builds a model in a step-by-step fashion as do other boosting approaches, which generalizes them by facilitating the optimization of an arbitrary differentiable loss function (Kannan & Vasanthi 2019; Chen et al. 2019). Trees are added to the ensemble model one at a time and are adapted to correct the prediction mistakes of prior models. This is a type of model of a machine learning ensemble known as boosting. For optimization of the arbitrary differentiable loss function and gradient descent, models fit with any algorithm. This gives the technique its name, "gradient boosting" as the gradient of loss is minimised as the model fits, much like a neural network.

Genetic Algorithm (GA)

GA follows the iterative learning theory, which was first introduced by Holland. This methodology operates on a principle close to that of natural-system genetic simulations (Figure 3.2. represents architecture of the GA technique). This algorithm initially used to identify individuals with a permanent population using a space snapshot. The role of exercise

is designed for individual evaluation. Any operations are carried out to develop new generations (Trivedi & Dey 2014).

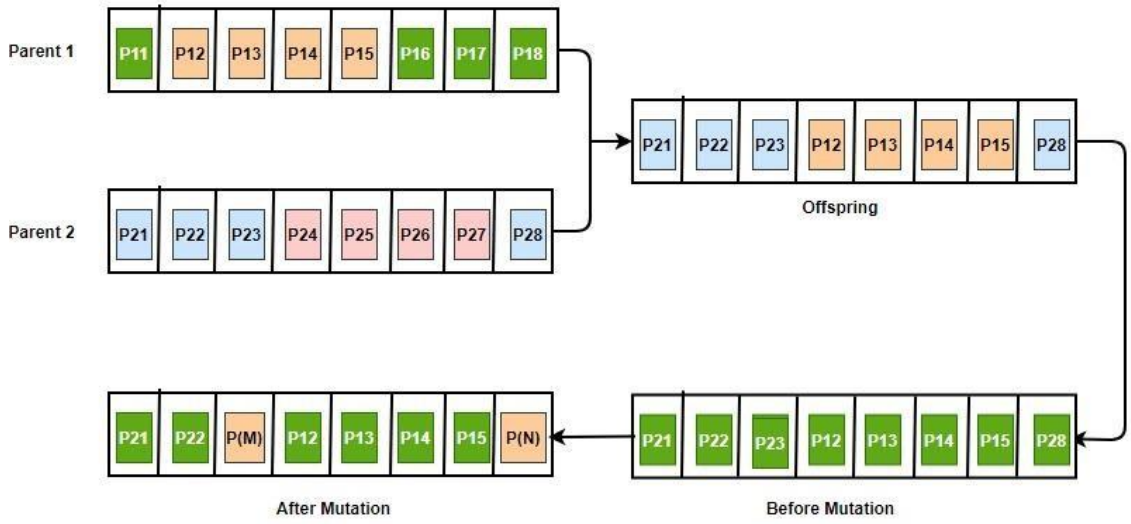


Figure 3.2. Architecture Explaining the GA Technique

Several studies have examined the use of GA in time-consuming tasks such as selecting features, which aim to select a specific small subset of features from the entire set of features (Xue et al. 2015).

AdaBoost

AdaBoost (Yuan & Ma 2012) produces a variety of sequential base classifiers as one of the typical learning algorithms using ensemble technique by changing the weights over training instances. In AdaBoost, instances misclassified by the present classification techniques are given a greater weight and vice versa. This allows classifiers to concentrate on instances which are not adequately categorized in previous classifiers and to create new base classifiers. As long as the base classifier is slightly more reliable than random inference, the upper limit of training error in AdaBoost typically decreases monotonically. The distinguishing attribute of the final determination shall always be weighted by each base classifier when combined.

Performance Metrics

Performance assessment of the proposed work is accomplished using the following measures. Confusion Matrix is utilized to evaluate the performance of a learning model. Four terms, related to the confusion matrix are applied to establish the performance matrices. The number of cervical cancer patients classified as cancer patients is True Positive (TP). False Positive (FP) is the number of non-cancer patients classified as cervical cancer patients. True Negative (TN) is the number of patients that are classified as non-cancer patients without cervical cancer. False negative (FN) is the number of patients classified as cancer patients without cervical cancer (Ray & Chaudhuri 2021).

Accuracy

It is the ratio between numbers of correctly predicted instances to the total number of instances.

$$\text{ACCURACY} = \frac{\text{TP+ TN}}{\text{TP+ FP+ FN+ TN}} \quad \ldots\ldots\ldots\ldots\ldots (3.6)$$

Precision

It evaluates the ratio of individuals predicted to be cancer patient and total number of cancer patients.

$$\text{PRECISION} = \frac{\text{TP}}{\text{TP+ FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

Sensitivity and Specificity

For all health care, diagnosis and procedures are imprecise and subject to error rates. The standard method for evaluating this medical error is a sensitivity and specificity test. A diagnosis and a medical examination should be differentiated. A test is one of many terms used to describe the medical condition; diagnosis is a mixture of numerous tests and observations that demonstrate the patient's pathophysiology. Sensitivity/specificity assessment is necessary for testing and diagnosis are shown in equation 8 and 9. There are test outcomes and an objective indicator of truth or theory in the medical dataset and analysis

of sensitivity and specificity. The consistency of test checks done by if-then rules governs the similarity of a new test result to the expected value. The best fit rule specifies the class composition of the study when defining an example of a test.

True Positive (TP) specifies the number of correct positive predictions (classifications); True Negative (TN) indicates the number of correct negative predictions; False Positive (FP) represents the number of incorrect positive predictions; False Negative (FN) identifies the number of incorrect negative predictions.

The two measures are:

$$\text{SENSITIVITY} = \text{RECALL} = \frac{TP}{TP + FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

$$\text{SPECIFICITY} = \frac{TN}{TN + FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

Sensitivity measures a test's tendency to be positive when the condition is present, or how many positive examples of tests are remembered. The sensitivity measures, in other words, test how often anyone finds what they are looking. It falls within several near-synonyms: false-negative rate, recall, type I error, type II error, omission error, or alternative hypothesis. Specificity tests a test's tendency to be negative if the condition does not exist, or how many negative test instances are omitted.

In other words, the specificity tests how often they are searching for what anyone can find. It falls within various near-synonyms: false-positive rate, accuracy, type I error, type II error, commission error, or null hypothesis. Predictive precision provides an overall assessment. Only findings that give high values for all three measures can be put with a high confidence level.

The AUC under the ROC curve is one of the most relevant parameters for assessing and rating the output efficiency of the classification and prediction models with a balanced sample condition; that is, the presence and the number of the absence of cases of HDs are roughly equal in the training and test data sets. However, if the data set contains unbalanced

samples, the F-score is the most critical parameter for the consistency evaluation classification and prediction models. In our present study, the AUC would be an appropriate way to rate the classification and prediction models' output levels.

F1-Score

It is the harmonic mean between precision and recall/sensitivity.

$$\text{F1-SCORE} = \frac{2(\text{PRECISION} \times \text{RECALL})}{\text{PRECISION} + \text{RECALL}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

AUC-ROC Curve

AUC-ROC or, AUROC (Area Under the Receiver Operating Characteristics) is a probability curve denoting the ability of a model to differentiate between classes in case of binary classification. ROC indicates exchange between True Positive Rate (TPR) and False Positive Rate (FPR). AUC signifies degree or measure of separability where the value closer to 1 means an algorithm effectively classifies patients with and without cancer.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.11)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.12)$$

Kappa Statistics

A chance-corrected method for evaluating agreement (rather than association) between raters is Cohen's kappa () statistic (Chalak et al. 2020). Kappa is defined as follows:

$$K_{STAT} = \frac{A_{OBS} - A_{EXP}}{N - A_{EXP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.13)$$

Where, A_{EXP} is the number of agreements predicted by chance, N is the total number of observations and A_{OBS} is the number of agreements observed between raters.

Chapter 4

System Design

Feature Selection (FS)

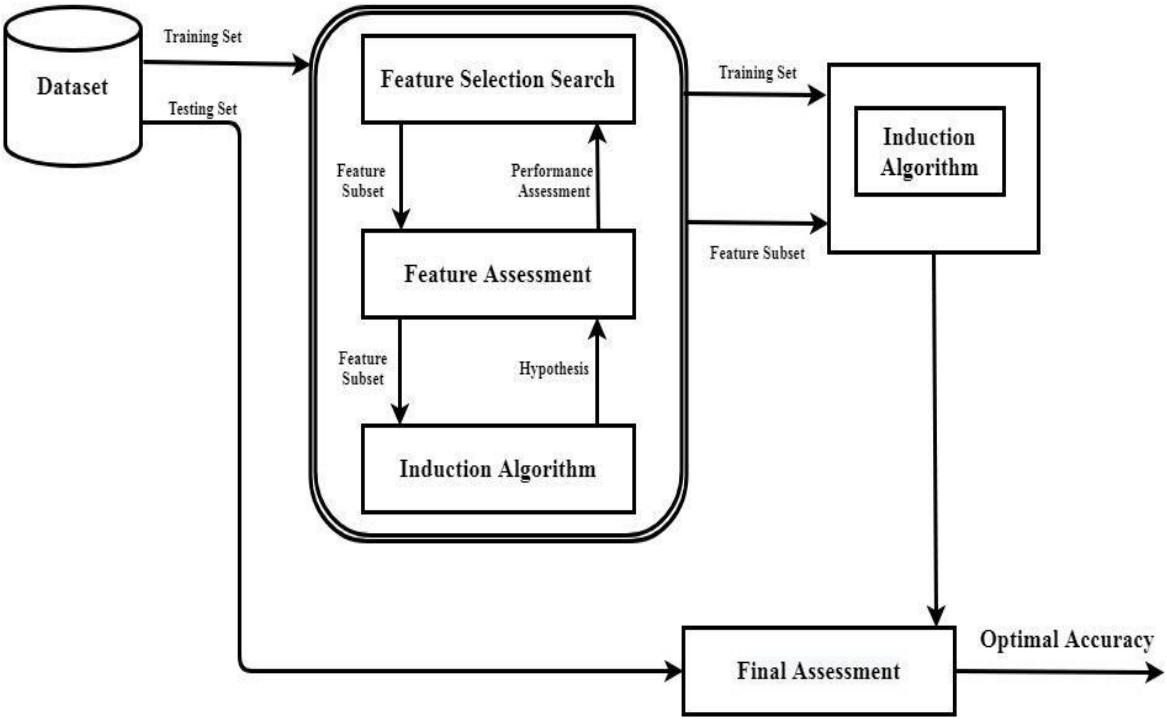


Figure 4.1. Wrapper Algorithm Approach for FS

Feature Selection (FS) is selecting a sub-set of original variables present in a dataset. Induction algorithms executed using that sub-set to generate a classifier with the highest possible accuracy. The significant point in FS is that it selects a set of variables from the existing collection of variables without creating any new variable i.e., without extracting any variables or constructing new variables (Kittler 1986; Rendell & Seshu 1990).

Degradation of performance or reduced prediction accuracy level with many unnecessary variables have noticed in practical ML techniques following top-down induction approach like DT (e.g., CART, ID3, and C4.5) algorithms and instance-based algorithms like instance-based learning (IBL) (Kohavi & John 1997).

Powerful algorithms to handle irrelevant variables like NB shows improved accuracy level at the initial stage, however, the performance may decrease rapidly after the addition of correlated variables, even when the variables are relevant (Kohavi & John 1997).

The FS techniques have several advantages e.g., minimization of cost to acquire data and interpretation of the models generated by classification methods become much easier (Wah et al. 2018). Commonly FS methods are subdivided into Wrapper, Filter, and Embedded methods (Ladha & Deepa 2011; Naqvi 2012). Different researchers discussed the usages, advantages, and disadvantages of Wrapper, Filter, and Embedded methods in their research studies (Ladha and Deepa 2011; Bolon-Canedo et al. 2014). Embedded and Wrapper methods generally interact with the classifier and depend on classification techniques whereas the Filter methods work independently and for that faster in nature. The main advantage of the Wrapper and Embedded method is a simple technique. At the same time, these two methods have some drawbacks like overfitting the model. Relief, chi-square, correlation-based feature selection, and information gain are some of the popular methods of Filter based methods. Whereas wrapper-based methods apply searching techniques and examples are sequential backward elimination, sequential forward selection, etc.

In the FSS technique, an algorithm is executing the problem of selecting the most relevant subset of variables from a dataset eliminating unnecessary variables by which the maximum accuracy is generated. In the wrapper algorithm approach, the FSS acts as a wrapper surrounding the induction algorithm. By using the induction algorithm itself, the FSS performs a search to extract an efficient subset of variables. In the wrapper approach, the induction algorithm was performed as a black-box (Kohavi & John 1997). The induction algorithm is used on the dataset, which has practically divided into internal training sets and holdout sets with different sets of variables removed from the dataset. The induction algorithm is executed on the final subset of variables generated from the variable subset that produces the highest assessment on which the induction algorithm is executed. The subsequent classifier has tested on a separate test set that has not been used for the search (Kohavi & John 1997). Wrapper algorithm approach for FS is shown in Figure 4.1..

Forward and Backward Greedy Algorithm

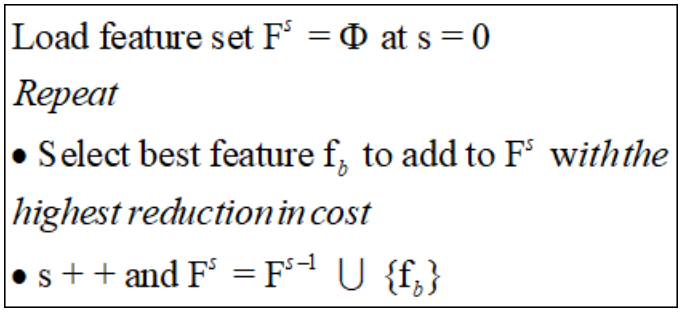


Figure 4.2. Forward Greedy Algorithm (Lee et al. 2011)

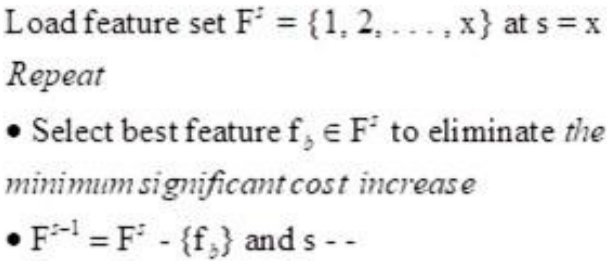


Figure 4.3. Backward Greedy Algorithm (Lee et al. 2011)

Sequential Forward Selection (SFS) (Panthong & Srivihok 2015) is the simplest greedy search algorithm. At the beginning SFS starts its operation with an empty set of features, it adds individual unused attribute from the available set iteratively. After each round the efficiency of the added attributes is evaluated using cross-validation. Only the attribute with maximum accuracy is added to the new set. After that the new iteration is started with modified set of features. The SFS algorithm therefore includes features that give the object function a largest value; Figure 4.2. illustrates a greedy forward algorithm. On the other hand, sequential backward selection (SBS) (Panthong & Srivihok 2015) performs in the reverse direction. It starts with complete set of features and removes individually the lowest performing features iteratively from the given set of attributes. The attribute with least decreasing performance is removed at last from the selection using cross-validation. The SBS is two way advantageous; firstly, it can remove several features and secondly it allows for backtracking, when a subset of features worsens the results produced by the previous

iteration, certain features previously removed can be included in the present re-evaluation subset. Figure 4.3. shows the backward greedy algorithm.

Feature Selection Methods

There are several methods for choosing features in the area of machine learning. The primary purpose of such approaches is to remove outdated or redundant features from the dataset. There are two types of FS methods, namely, the wrapper and filter method. The wrapper checks and selects attributes based upon the target learning algorithm's accuracy estimates. The wrapper effectively checks function space by omitting some characteristics and checking the effect of the elements' omission on the predictive metrics. The position that makes a significant difference in the learning process means that it is essential and is viewed as a high-quality function.

On the other hand, whatever the learning algorithm, "Filter" uses the data's general features and functions. The number correlation between a set of characteristics and the target function is explicitly used by "Filter." The target variable value determines the amount of similarity between the target variable and the characteristics. Filter-based solutions are not classification-based and are typically quicker and more flexible than wrapper-based methods. We also have a low complexity of computation. In filter algorithms, the features are first rated and graded by classmark significance and then picked by a threshold value (Witten & Frank 2002). In each of these applications, collection algorithms have an assessment value for every feature. This paper uses four significant filter approaches: Information-gain, Relief-F, One-R, and Gain-ratio. The following sections describe the characteristics of these approaches.

Information Gain

Information Gain is one of the methods commonly used in a variety of applications for evaluating features. This approach controls all functions according to a user-defined target value. The entropy description for the feature rank is embedded in the cycle of Information

Gain. This approach was formulated to estimate each attribute's quality using entropy by calculating the difference between the prior entropy and the post entropy (Kononenko 1994).

Information gain (relative entropy or Kullback-Leibler divergence) is a function of the difference between two probability distributions in probability theory and information theory. This approach assesses a feature M by measuring the amount of information obtained from factor N of class (or group), defined as follows:

$$I(M) = H\left(P(N)\right) - H\left(P\left(\frac{N}{M}\right)\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

In particular, it tests the difference between the marginal distribution of measurable N on the assumption that it is independent of $H(P(N))$ and the conditional distribution of N on the assumption that it is dependent on $H(P(\frac{N}{M}))$. If M is not expressed differently, N will be independent of M, which means that M will have limited benefit value for data and vice versa.

Relief-F

Kira & Rendell (1992) developed Relief using the distance-based metric method, which weighs each feature depending on its significance (correlation) to the target class. However, reliability is inefficient as it can handle only two-class problems and does not handle redundant features. The updated version of Relief, known as Relief-F (Kononenko 1994), can manage multi-class problems and handle incomplete and noisy data sets. Nonetheless, the elimination of redundant functions fails. The selection of features is an instance-based approach that evaluates a function by extracting samples from different but similar classes and their output. For each feature M of the same class and each of the different classes, Relief-F selects a random sample N of its closest neighbours. Kononenko (1994) defines M as the number of weighted variations and the same classification between different classes.

One-R

One-R is a simple algorithm proposed by Holte. It produces one rule in the training data for each attribute and then selects the rule with the least error. This method considers all numerically valued features continuous and uses a straightforward approach to split the range of values into several intervals of disjoint. It handles missing values by marking a specific attribute as "missing." This system is among the oldest of these. It produces simple, one-feature rules. Although it is a type of minimal classification, it may be useful as a benchmark for other learning schemes to determine a baseline performance (Yildirim 2015).

4.3.4 Gain ratio

Gain ratio (GR) is a shift that reduces the bias in information gain. The GR will take into account the number and size of divisions when selecting an attribute. It corrects information gain by recognizing the inherent knowledge of a split. The intrinsic information is entropy of branch-wide instance distribution (i.e., how much information we need to say what branch an instance belongs to). The attribute value decreases as the information intrinsically increases (Karegowda et al. 2010). Equation 4.2 gives the way gain-ratio of an attribute is calculated.

$$GR(\text{attribute}) = \frac{\text{Gain}(\text{attribute})}{\text{Intrinsic-Info}(\text{attribute})} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

Feature Subset Selection

The process of selecting a subset of suitable features (variables, predictors) for use in machine learning and statistics model construction is the feature selection, also known as variable selection, selection of attributes, or selection of variable subsets. Methods for choosing the roles are used for various purposes:

- Simplifying models to make them easier for researchers/users to understand (Luo et al. 2018)
- Short time to train

- To escape the dimensionality curse
- Better generalization by decreased overfitting (formally, reduction of variance (Luo et al. 2018))

The central premise when using a selection technique for features is that the data includes certain features that are either obsolete or outdated and can therefore be eliminated without much information loss. Two distinct notions are interchangeable and irrelevant, because one relevant feature may be redundant in the presence of another relevant feature to which it is strongly correlated.

It is necessary to distinguish techniques for selecting features from extracted features. Function extraction creates new features from the original feature sets, and feature selection returns a subset of features. Feature selection techniques are often used in multi-function domains with relatively few samples (or data points).

Searching the Feature Subset

Subset selection checks the appropriateness as a category of a subset of functions. Subset selection algorithms can be broken up into wrappers, filters, and embedded methods. To check for possible feature space, wrappers use a search algorithm and evaluate each sub-set by running a model on the subset. Wrappers can be costly in computational terms and risk being over-fitted to design.

Filters in the search approach are similar to wrappers but instead of evaluating against a model a simpler filter is evaluated. Embedded techniques are incorporated and specific to a model. In the following section, greedy, ranked, and best first search methods are elaborated.

Greedy Hill-Climbing Search (GS)

If a feature selection algorithm is to work on data with a large number of features, it is important to check the space of the feature subsets within reasonable time constraints. One simple search technique, called greedy hill-climbing, takes into account local changes to the current subset of features. The most common method to search using greedy hill-climbing,

which iteratively evaluates a candidate subset of features, then modifies the subset and assesses whether the new subset is an improvement over the old one.

Assessing the subsets requires a scoring metric which grades a subset of features. Local modification often involves adding or removing a single characteristic from the subset. If the algorithm only considers additions to the subset of features it is referred to as the forward selection and the deletions are referred to as the backward elimination method (Chen et al. 2019). The alternative approach, called step-wise bidirectional search, is using both the addition and the deletion.

It includes each of these variations; the search method will take into account all possible local changes to the current subset, and then select the best or first adjustment that will boost the value of the current subset of functions. In both cases, it is never rethought after a change has been detected. The stop criterion varies by algorithm; possible conditions include: a sub-set score reaches a threshold; the maximum allowable run time has been exceeded by a program etc.

Best-First Search

Best-first search is a search algorithm exploring a diagram by expanding the most promising node chosen according to a given law. Judea Pearl defined the best-first search as estimating the promise of node n by a 'heuristic evaluation function' which can, in general, depend on the definition of n, the description of the aim, the information gathered by the search up to that point, and, most importantly, any additional knowledge of the problem domain.

Some scholars have used "best-first search" to explicitly refer to a heuristic search that aims to determine how close the end of a path is to a solution so that paths that are considered to be closer to a solution are first extended. A particular type of search is called greedy best-first search or pure heuristic search. The efficient selection of the current best extension candidate is typically done using a priority queue (Cenamor 2017). The Best First Search is an important AI search technique that enables the search route to be backtracked. Like greedy hill-climbing, the best first step through the search space by making local changes to the

current subset of features. The main difference of Best First with hill climbing is that, if the route being followed begins to look less promising, the Best First method will back-track to a more promising previous subset and continue searching from there. For a specified time, the best first search will explore the whole search space, so it's common to use a stop criterion. It normally involves limiting the number of completely expanded subsets which lead to no improvement.

Rank-Based Algorithms

The feature selection algorithms that generate an ordered list of features that indicate the relative value of each feature based on a particular condition are known as Ranker algorithms. The algorithm analyzes every attribute and produces a rank-list of attributes. Algorithms for Ranker can be classified into two categories: Filter and Wrapper. The rationale behind the filter methods is based on mathematical measurements such as information theory, probability estimation, probability calculation, error measurement, etc. Methods based on information theory such as information gain, gain ratio, etc., concentrate on covering maximum information, whereas methods based on likelihood like relief and correlation are likely to focus on class consistency over a given value of a function. To obtain a subset from ranker methods, a threshold on the number of features to be selected from the rank-list is necessary. The characteristics covered by the threshold will be divided, thereby obtaining a dimensionally reduced set of data. The characteristics covered by the threshold will be divided, thereby obtaining a dimensionally reduced set of data. There is no standardized procedure on function count to determine the threshold. This is because the accuracy varies with several factors including the nature of the ranker algorithm, the data skew in the data set in question, and the nature and characteristics of the data set (Zelenkov, Fedorova & Chekrizov 2017; Mafarja & Sabar 2018).

Chapter 5

Application of Data Mining Techniques for Avoiding Underestimation of An Event

Introduction

Data mining techniques (DMTs) have often been used to extract useful information from massive data sets and patterns (Liao et al. 2012). There is evidence of widespread use of DMT in the diagnosis of diseases. However, no single method has shown consistent outcomes, and thus, researchers have proposed a hybrid approach (Jothi & Husain 2015). In most cases, the results suffer from the over-fitting or under-fitting of data affecting predictions. Incorrect treatments are irreversible and have long term impacts. There are instances of missed treatment (Type 1 error) or treating the wrong ones (Type 1 error). The individual methods showed varying levels of accuracy, and diagnosis with the highest accuracy levels, say 75% or so, is used for suggesting treatments.

In some cases, the accuracy can be around 90% or more. Thus, the question arises as to how can the prediction levels be enhanced with existing methods? The authors aim to ensemble DMTs in this paper to explore the possibility of increasing the accuracy of the prediction. Two databases - patient and treatment database were used to identify the occurrence of cervical cancer. Three supervised learning methods - decision tree (DT), random forest (RF), and logistics regression, have been used to identify the importance of variables. The prediction was carried using original datasets, and revised datasets comprising variables found significant. A comparison of the outcomes across two datasets was made to conclude the cause of the disease.

Several kinds of literature (detailed in the next section) on cervical cancer point out that HPV OF certain types leads to warts in patients, but all warts do not show signs of cancer either in the short term or in the long run. This uncertainty is so because initial detection of symptoms (such as the formation of warts) does not mark cancer in a patient, as this is temporal. Warts so observed may, after some time, turn out to be carcinoid. There are instances where warts prevalent in a human body even after six months or so may not require cancer treatment, whereas a 3-month-old wart may turn out to be cancerous. Thus, there are possibilities of

incorrect estimation of HPV in many cases due to limitations in the study of this disease. In such cases, analysis requires mining the patients' data set (with warts) diagnosed with or without cancer and patient (with warts) who underwent treatment. Unfortunately, there is a lack of a centralized database to arrive at the right diagnosis. Different databases, one about patients detected with warts and other - patient (with warts) undergoing treatment, are available. There are few attempts to derive association rules combining patient characteristics and treatment response for firming up the right diagnostics.

In this article, diagnosis of information available in the World-Wide-Web was carried out to develop the information hierarchy (of cervical cancer) to facilitate analysis. In this paper, the authors aim to combine data mining methods to explore the possibility of increasing the accuracy of the forecast.

The paper has been organized into seven sections. The next section gives a brief overview of the disease - cervical cancer and the Human papillomavirus virus (HPV); section 4 discusses the different DMTs used in this study; section describes the dataset, and section 6 analyses the data. Section 7 discusses the results of the analyses, and section 8 concludes the research work.

Human Papilloma Virus (HPV)

General Description

The human papilloma virus (HPV) is the most widely recognized sexually transmitted viral infection. Sexually dynamic individuals are most prone to this infection (Koutsky 1997; Gabbey & Jacquelyn 2017). The disease's primary issue is the time gap between the actual time of infection and the time the virus gets manifested, following which the treatment begins. This time gap is because warts that develop due to HPV infection has no signs or side effects in the infected person. As a result, the disease unconsciously passes on to their sexual accomplices. In the long run, the infection causes cervical cancer.

Developing countries witness women's maximum death on account of cervical cancer while worldwide, it ranks second as a terminal disease. Study shows that sexual transmission of

human papilloma virus (HPV) causes cervical Intra-epithelial neoplasia and invasive cervical cancer (Gabbey & Jacquelyn 2017). Every year 5,10,00 new cases and 2,88,000 deaths are reported worldwide (Shouman et al. 2012). The uniqueness lies in the fact this particular malignancy is detected at the productive age of women. The disease gets detected at an average age 30 to 34 years, and the tops at the age of 55 to 65. The mid-age being 38. On average, sexually dynamic women get infected with genital HPV by 50 years of age (Sankaranarayanan & Ferlay 2006).

India accounts for around one-third of global deaths due to cervical cancer (http://www.who.int/hpvcentre) against 6.6% of global infection. The primary types include serotypes 16 and 18, accounting for 76.6% of cases. The women developing warts account for around 2–25% of sexually communicated diseases (http://www.who.int/hpvcentre).

Cervical cancer is detected using the Pap test followed by colposcopy test. The serious problem lies in the fact that there is no recommended conventional treatment for HPV infection. Doctors resort to wait-watch on the symptoms to convert to the pre-cancerous stage, i.e., observe for wart and or pre-cancerous changes of the cervix. In this paper, the authors propose to apply data mining tools to derive meaningful insights for the treatment of cervical cancer.

Cervical Cancer – Profile Analysis

Profile analysis serves as a convenient way to represent information and has been used by researchers to describe information about groups and families of sequences. Some authors (Gribskov et al. 1987) used this approach for the study of proteins. HPV virus-infected persons develop warts in the throat or genitals and can cause cancer in these regions or in the head and neck. Warts appear benign in the initial stage but later manifest to be malign, and the significant problem is that cancer is in a later phase of development at this stage. Patients who undergo periodic Pap tests may have early detection of the disease. These findings can improve perspective and increase chances of survival (Gabbey & Jacquelyn 2017).

HPV types 16 and 18 are the cause of about 70% of all cervical cancer cases worldwide. HPV antibodies that forestall HPV 16 and 18 infections are currently accessible and can decrease the incidence of cervical and other anogenital cancers.

A report citing the different perspective of this disease provides HPV related insights; and mentions factors adding to cervical cancer; cervical cancer screening rehearses; HPV antibody presentation; and other significant vaccination pointers. Study shows underestimations of the prevalence of HPV in many cases due to limitations in methods of research of this disease (Schmitt et al. 2010). The chance of this infection increases with the severity of the lesion. Cervical lesions between 41% and 67% of high grade and 16%-32% of low grade contribute to 70% of cases, namely HPV 16 and 18 (Shouman et al. 2012). 20% of other cervical cancer types include the HPV types – 31, 33, 35, 45, 52, and 58.

The most challenging aspect of this infection is that in many cases, HPV goes away on its own, so there is no treatment required. Instead, the doctor would preferably want the patient to go for repeated testing on a half-yearly or yearly basis to check for the infection's persistence. In such a case, six months may be too late for the treatment.

The US Food and Drug Administration (FDA) accepted the principal DNA test for HPV in 2014. Updated rules suggest that women have their first Pap test, or Pap smear, at age 21 and be tested for HPV simultaneously, paying little mind to the beginning of sexual activity (Gabbey & Jacquelyn 2017). After that, women between the ages of 21 to 29 ought to have a Pap test at regular intervals. Standard Pap tests help recognize anomalies in cell structure, which serves as the caution against cancer growth or any such severity.

The doctors follow the general convention of screening women in the age group of 30 to 65 at an interval of five years based on Pap and HPV diagnosis. For age less than 30 years, such tests are prescribed if the smear test shows inconsistent results. For example, if patients have any of the 15 HPV strains, the likelihood of cancer is higher. The frequency of screening, in such cases, increases as in many cases, it takes ten years to get malignant. Such FDA affirmed tests for men are not present (Gabbey & Jacquelyn 2017).

The periodicity of the test, so far, recommended may have high variation depending on a host of factors. The availability of patient data and the application of data mining techniques can lead to a more precise recommendation of the test's periodicity and predict the probability of cancer occurrence from warts detected.

Web Information Diagnosis

The authors carried out a diagnosis of information available in the World-Wide-Web using Sem-Rush software. The result highlightstwo aspects – age and HPV type 16; besides culls out various variants associated with age, types, cancer, warts, gender, sex, and related facets.

Information Hierarchy of Cervical Cancer

Figure 5.1. shows the ontology of cervical cancer developed by the author, from the study of literature and web-information analysis.

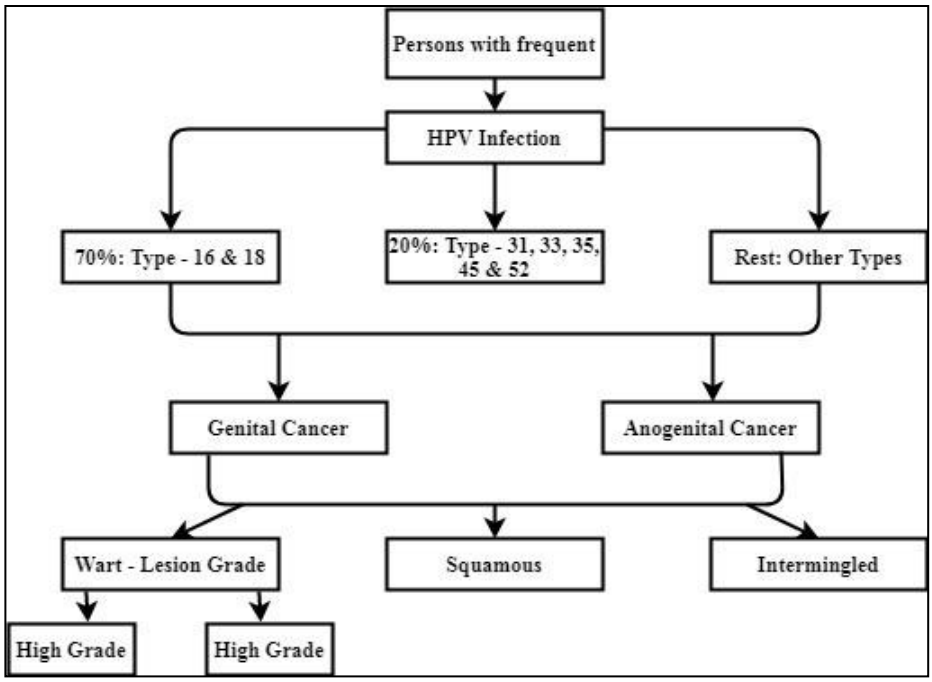


Figure 5.1. Profile of Cervical Cancer

The research question is, therefore, how to minimize type 1 and type 2 errors in the treatment of cervical cancer. In other words, the sub-questions are:

i. What are the major causes of this disease?
ii. When to begin treatment after detection of warts or abnormality in Pap test?
iii. What is the probability of correct diagnosis given the age, gender, Time elapsed before treatment (month), the number of warts, types of wart (Count), Surface area of warts(mm2) and other factors?

Data Mining in the Analysis of Diseases

Cervical cancer mostly affects women with 5,70,000 new cases in 2018. The majority of these cases are from developing countries. The detection of cancer at the right time, effective screening and treatment programmes can make the lives of cancer patients better (https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/). The limited availability of resources screening and diagnosis from a Computer-Aided Diagnosis (CAD) point of view becomes difficult. Also, sometimes, patients do not participate in routine testings. Thus, the correct, timely diagnosis and estimation of specific risks of each patient are an assurance of proper treatment. The experience of the doctor and their abstract choice affect the results in a large portion of these screening techniques (Singh 2005). A survey could be applied to patients to decide the riskiest groups and to reduce pointless screenings. Hence the patients can resort totesting based on the likelihood of cancer.

There are several references to the use of different DMT to discover numerous investigations about the assessment of some therapeutic screenings. Sen et al. (2013) in their paper have introduced an artificial neural network in pancreatic ailment diagnosis dependent on a set of symptoms. The authors found thatdetection using neural network had higher accuracy than other manual methods (Gabbey & Jacquelyn 2017). Fernandes et al. displayed a regularization-based TL way to deal with exchange the contribution type for each component on direct models. Kalyankar & Chopde (2013) demonstrated the use of Genetic Algorithms, Artificial Neural Network, Hierarchical Clustering, Neuro-Fuzzy framework, Raman Spectra. To arrive at an overview of various kind of cancer, for example, skin cancer, bosom cancer, pancreatic cancer. In this study, classification performance varied in the range of 80.5% and 95.8% depending upon the cancer type and grouping. In 2016, Kanimozhi &

Karthikeyan (2016) showed the use of mining procedures in the investigation of various coronary illness. The authors utilized different databases and found that the achievement rates have fluctuated between 45% to 99.1% contingent upon arrangement strategies and traits in the different databases. Fatima & Pasha (2017) exhibited similar examination of various machine learning calculations for diagnosis of various maladies, for example, coronary illness, diabetes infection, liver sickness, dengue ailment, and hepatitis ailment. They demonstrated that different techniques showed different levels of accuracy for a particular disease, while the best method for a specific ailment did not perform equally for a separate category of illness or a database. Sankaranarayanan & Ferlay (2006) showed that Naive Bayes had an accuracy of 97% while Neural Network demonstrated 70 percent achievement rates. Thus, the diagnosis based on the best result may lead to the occurrence of both type 1 and 2 errors. This issue gets resolved by integrating the techniques and identifying a numberof association rules derived from the disjoint set of the DMT. The researchers have so far not used this approach in diagnosing diseases, especially cervical cancer.

Datasets

Table 5.1. depicts the attributes in the dataset obtained from the University of California, Irvine – the data related to patients in Hospital Universitario de Caracas in Caracas, Venezuela. It captures 35 characteristics of cervical cancer of 858 patients (UCI 2019). The attributes comprise information associatedwith demography, habits, and historical medical records (UCI 2019).

Table 5.1. Attribute Information of First Dataset

Feature	Type	Feature	Type
Age	Int	STDs: pelvic inflammatory disease	bool
# of partners	Int	STDs: genital herpes	bool
Age of 1st intercourse	Int	STDs: molluscumcontagiosum	bool
# of pregnancies	Int	STDs: AIDS	bool
Smokes	bool	STDs: HIV	bool
Smokes: years	Int	STDs: Hepatitis B	bool
Smokes: packs/year	Int	STDs: HPV	bool
Hormonal Contraceptives	bool	STDs: Number of diagnosis	Int
Hormonal Contraceptives years	Int	STDs: Time since first diagnosis	Int
IUD	bool	STDs: Time since last diagnosis	Int
IUD years	Int	Dx: Cancer	bool
STDs	bool	Dx: CIN	bool
STDs number	Int	Dx: HPV	bool
STDs: condylomatosis	bool	Dx	bool
STDs: cervicalcondylomatosis	bool	Hinselmann: target variable	bool
STDs: vaginalcondylomatosis	bool	Schiller: target variable	bool
STDs: vulvo-perinealcondylomatosis	bool	Cytology: target variable	bool
STDs: syphilis	bool	Biopsy: class or target variable	bool

The second dataset relates to the treatment of patients with common wart types, sourced from dermatology clinic of Ghaem Hospital in Mashhad during the period January 2013 to February 2015. Table 5.2. presents the dataset that captures eight features of patients who underwent treatment using the immunotherapy method. The class attribute in these datasets is the Response to Treatment feature.

Table 5.2. Attribute Information of the Second Dataset

Feature name	Values	Mean ± SD
Response to treatment	Yes or No	
Gender	41 Man, 49 Woman	
Age (year)	15 – 56	31.04 ± 12.23
Time elapsed before treatment (month)	0 – 12	7.23 ± 3.10
The number of warts	1 – 19	6.14 ± 4.2
Types of the wart (Count)	1 – Common (47), 2 – Plantar (22), 3 – Both (21)	
Surface area of the warts (mm2)	4 – 750	85.83 ± 131.73

Variable Clustering and Importance

The authors used K-meansfor identifying the clusters. They used Decision Tree (DT) and Random Forest (RF) techniquesto identify importance of variables suggested by these methods respectively. The authors used Logistics-regression (LR) on the orginaland reviseddata set to predict the outcomes and compared the results to see the change in accuracy of prediction. The revised set consisted of data on important variables obtained from RF and DT analysis respectively.

K – Means Clustering

Analysis of the Disease Dataset: K – means analysis of the disease data set (Table

5.3. – Appendix A) show that significant variables include Age, First sexual intercourse, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives (years), IUD (years), STDs: vaginalcondylomatosis, STDs: Time since the first diagnosis, and STDs: Time since the last diagnosis. Table 5.5. shows the number of cases in each cluster.

Analysis of the Treatment Dataset: The K-means analysis (Table 5.4.) of dataset related to the treatment of patients indicate that the cluster where 76 out of 90 cases clustered indicate that number of warts (6) in women patients of age 32 years showed signs of cancer

in around seven months. However, there are four cases where cancer was detected in approximatelysix months, even though the number of warts is low (2 numbers). There is one case where cancer got detected in three months after detection of warts.

Table 5.4. K-Means Analysis of Dataset Related to Treatment

	Cluster				
	1	**2**	**3**	**4**	**5**
Sex	2	1	2	2	1
Age	32	15	26	32	35
Time	6.25	3.00	8.29	7.17	9.25
Number_of_Warts	8	2	4	6	4
Type	2	3	1	2	1
Area	350	900	177	49	504
induration_diameter	6	70	11	15	6
Result_of_Treatment	1	1	1	1	1

Table 5.5. Number of Cases in Each Cluster

Cluster	**1**	4.000
	2	1.000
	3	6.000
	4	76.000
	5	3.000
Valid		90.000
Missing		.000

Random Forest (RF) Analysis

RF analysis on original desease dataset shows Age, Number of sexual partners, Hormonal Contraceptives (years), IUD (years), STDs: Time since the first diagnosis, Dx: CIN and Dx: HPVas relatively important variables.

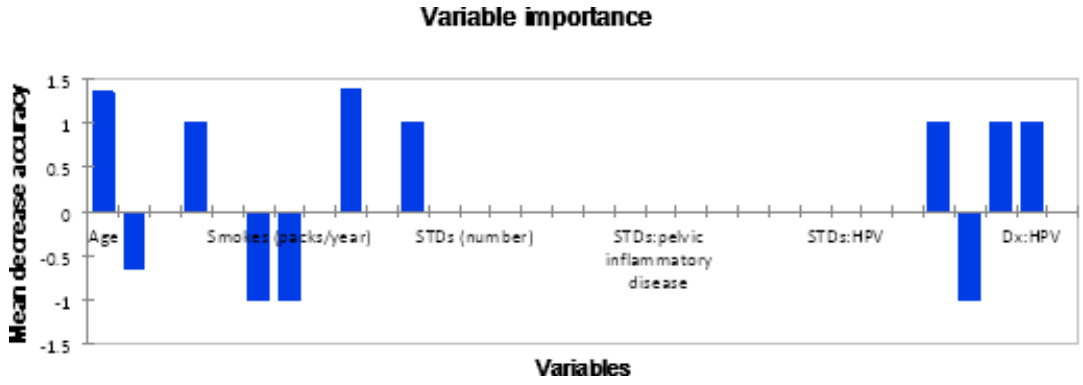


Figure 5.2. Random Forest Analysis of the Data on Disease Dataset

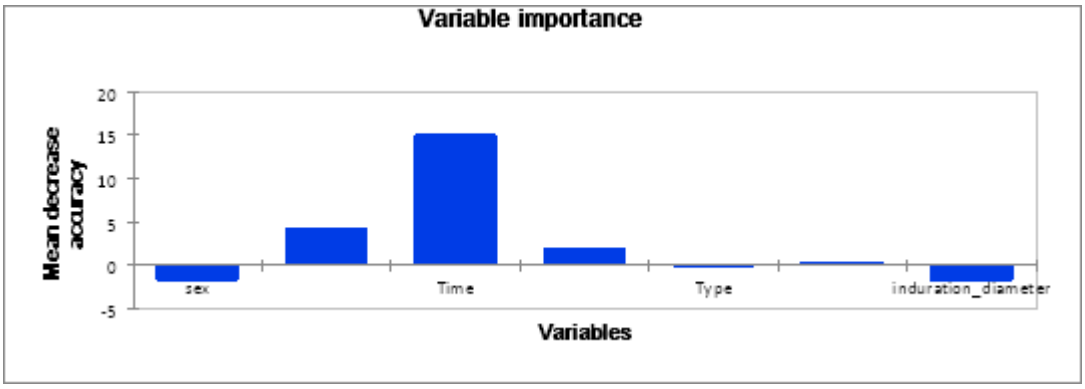


Figure 5.3. Random Forest Analysis of the Data on Treatment Dataset

Similar exercise on treatment dataset predicts Time, age, Number_of_Warts, and Area as relatively significant ones.

Decision Tree

Disease Dataset

The result from DT analysis of disease dataset is given in Figure 5.4..

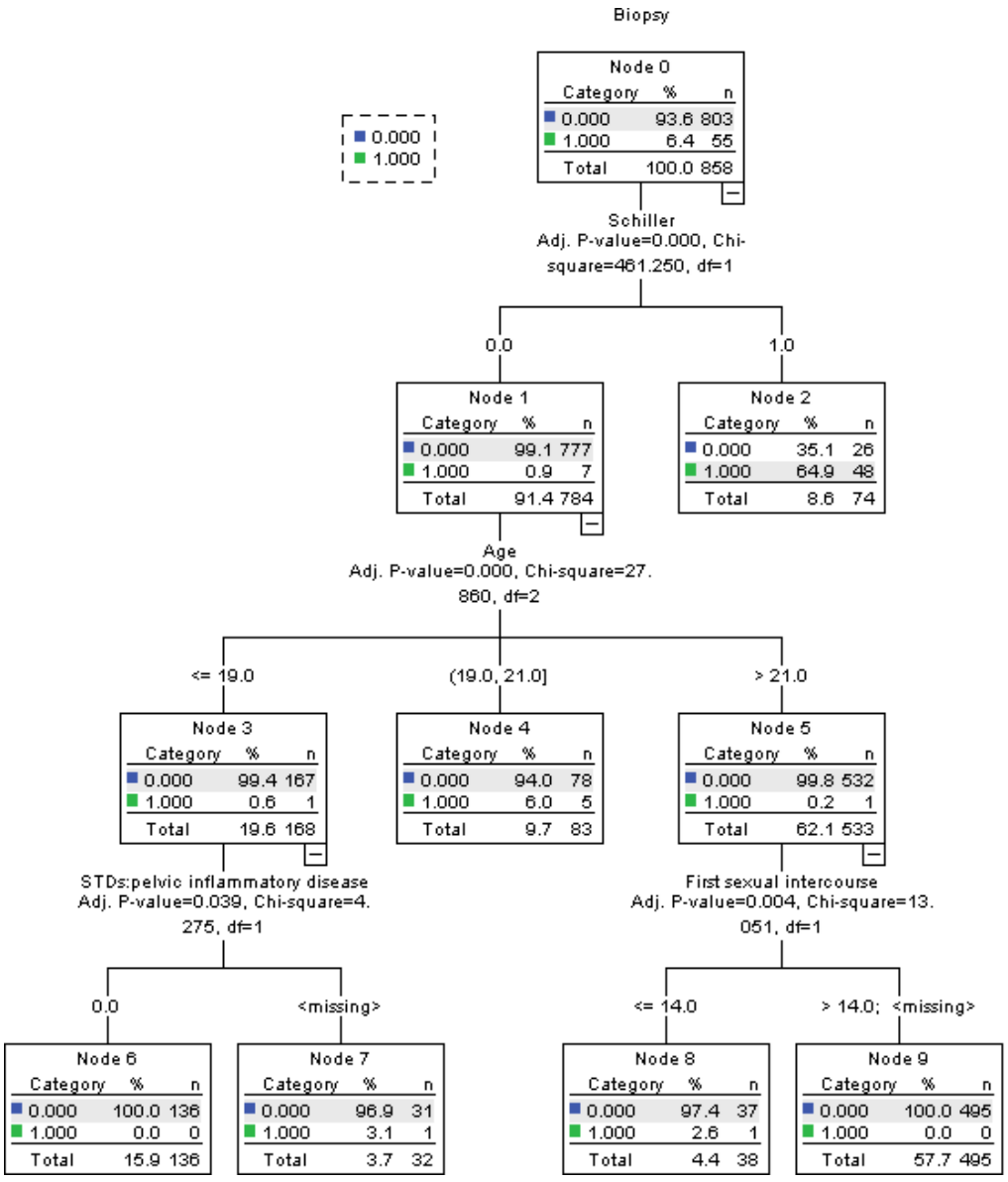


Figure 5.4. Outcome of the Decision Tree Analysis Carried Out on all the Variables of Disease Dataset

The result suggests the association rules as:

Schiller's test result is the significant factor for detection of cervical cancer.
For the Schiller's test result = 1, 64.9 % of women have cervical cancer.
For the Schiller's test result = 0, next best predictor is age
If age > 19 and <= 21 then, 6 % of women have cervical cancer
For the age <= 19, next best predictor is STDs : pelvic inflammatory disease
For STDs : pelvic = 0.0, 0.0 % women are attacked with cervical cancer
For STDs : pelvic = missing, 3.1% of women have cervical cancer
For the age > 21, next best predictor is First sexual intercourse
For First sexual intercourse <= 14, 2.6% of women have cervical cancer
For First sexual intercourse > 14, 0.0% of women have cervical cancer

The accuracy of findings is to the tune of 96.2% as shown in Table 5.6..

Table 5.6. Accuracy Level of Decision Tree Analysis

Classification			
	Predicted		
Observed	**0**	**1**	**Percent Correct**
0	777	26	96.8%
1	7	48	87.3%
Overall Percentage	91.4%	8.6%	96.2%
Growing Method: CHAID **Dependent Variable:** Biopsy			

DT analysis shows that variables Schiller Test, STDs: pelvic inflammatory disease and First sexual intercourse) are important and differ results of Random forest analysis excepting Age being common in both the findings. The excercise when further extended on revised data set comprising only variables found important from RF test will provide the relative accuracy levels.

Treatment Dataset

The decision tree with all the variables of treatment dataset predicts the association rules as follows.

For the time <= 5.250, 84.6 % of women got treated.

Time > 5.250 and <= 8.000 then, 100% of women were treated

If Time > 8.000 and <= 10.500 then, 83.3% of women were treated

For the time > 10.500, 29.4 % of women got treated.

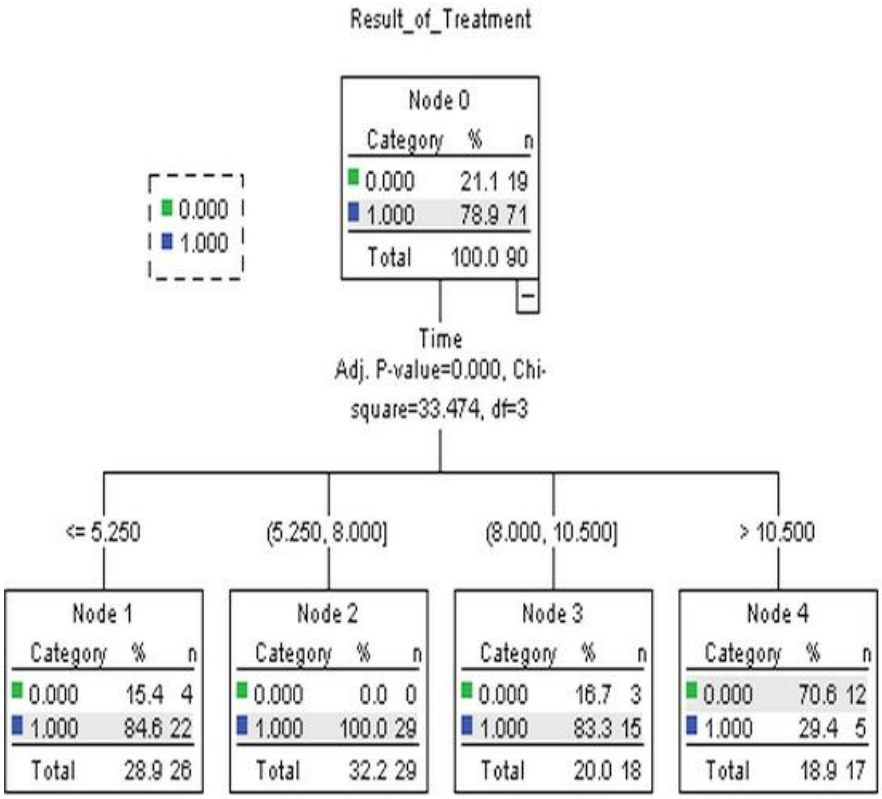


Figure 5.5. Outcome of the Decision Tree Analysis Carried Out on all the Variables of Treatment Dataset

Table 5.7. shows the extent of accuracy. The decision tree analysis on all variables predicts the accuracy of 86.7%. DT application on treatment dataset shows that time is the only important variable and thus differ from results ofRF analysis that predicts time, age, number of warts and area as significant. Time is common finding from both the analysis. The excercise when further extended on revised treatment-dataset comprising only variables found important from RF test will provide the relative accuracy levels.

Table 5.7. Prediction Accuracy

Classification			
	Predicted		
Observed	**0**	**1**	**Percent Correct**
0	12	7	63.2%
1	5	66	93.0%
Overall Percentage	18.9%	81.1%	86.7%

Growing Method: CHAID
Dependent Variable: Result_of_Treatment

Decision Tree Analysis on Relatively Important Variables Determined from RF Analysis

Disease Dataset

DT analysis on variables identified as relatively significant using RF analysis shows that Dx: HPV is the critical factor relating to the disease (Figure 5.6.). The approach predicted with 93.6% accuracy (Table 5.8.). However, it failed to predict the occurrence of illness using this approach. The cause being the lesser number of data points (6.4%) relating to patients with biopsy equal to 1. This method correctly predicted the instances of HPV patients who did not have cancer, i.e., with no type 2 error. This approach proved inferior to the results of DT analysis considering all variables in the data set.

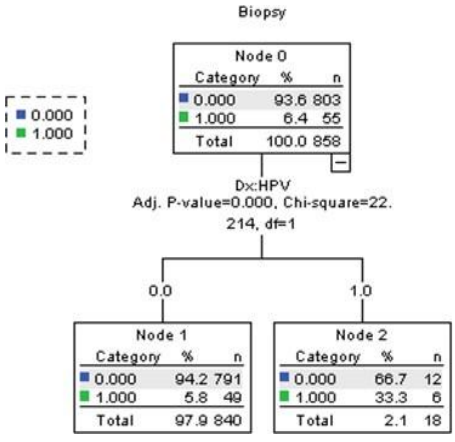


Figure 5.6. Outcome of the Decision Tree Analysis Carried Out on all the Relatively Important Variables from Disease Dataset

Table 5.8. Prediction Accuracy

Classification			
	Predicted		
Observed	**0**	**1**	**Percent Correct**
0	803	0	100.0%
1	55	0	0.0%
Overall Percentage	100.0%	0.0%	93.6%

Growing Method: CHAID
Dependent Variable: Biopsy

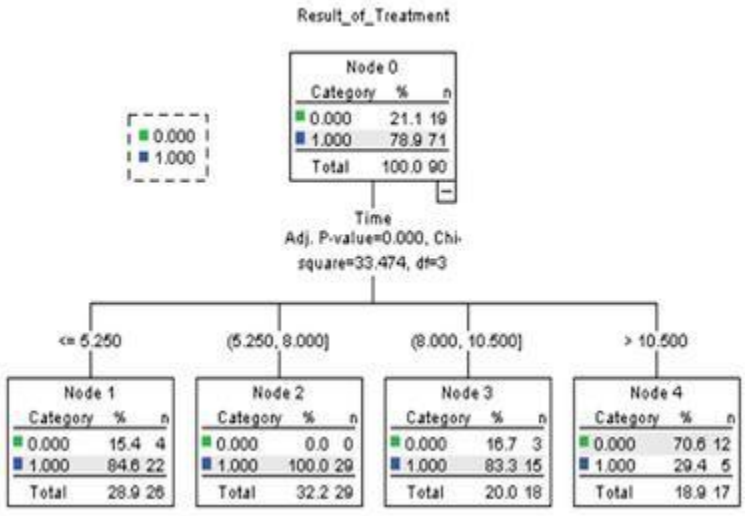


Figure 5.7. Outcome of the Decision Tree Analysis Carried Out on all the Relatively Important Variables from the Treatment Dataset

Treatment Dataset

Decision tree analysis on variables determined as relatively important using RF analysis shows that "Time" is the critical factor relating to the treatment of the disease (Figure 5.7.). The approach predicted with 86.7% accuracy (Table 5.9.). The occurrence of cancer had a prediction accuracy of 93% as against 63.2% accuracy of HPV patients without cancer (biopsy=0). Thus, the type 2 error was found to be higher than the type 1 error. The results from this approach did not differ from the Decision tree analysis carried out, taking all

variable in the treatment data set. In this data set the two categories of classification comprised data records in the ratio of 79:21 (Biopsy = 1: Biopsy = 0) percentage.

Table 5.9. Prediction Accuracy

Classification			
Observed	**Predicted**		
	0	**1**	**Percent Correct**
0	12	7	63.2%
1	5	66	93.0%
Overall Percentage	18.9%	81.1%	86.7%

Growing Method: CHAID
Dependent Variable: Result_of_Treatment

Logistics Regression

Disease Dataset

The Logistics regression on variables of disease dataset shows that noneof the variables are significant in predicting the disease.

Treatment Dataset

The Logistics regression on variables of disease dataset shows that Time is the significant variable along with the constant value. Equation 5.1 enables the prediction of the disease with Time as a significant variable.

$$p = \frac{1}{1 + e^{-(5.151 - 0.353 \times \text{Time})}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5.1)$$

The results showed an accuracy of 85.6% (Table 5.10.).

Table 5.10. Classification Table

Classification Table[a]					
Observed			Predicted		
			Result_of_Treatment		Percentage Correct
			0	1	
Step 1	Result_of_Treatment	0	6	13	31.6
		1	0	71	100.0
	Overall Percentage				85.6

a. The cut value is .500

LR Analysis of Revised Dataset Comprising Important Variables Determined from DT Analysis

In this approach, as well, none of the variables were found significant. The results obtained from Decision tree analysis (on the complete data set) proved to be the best so far with an accuracy level of 96.2%.

LR Analysis of Revised Dataset Comprising Important Variables Determined from RF Analysis

The variable – Time was found to be significant along with the constant. The findings appear similar to the LR analysis of the complete dataset. However, the accuracy level (Table 5.11.) was found to be higher in this approach.

Table 5.11. Classification Table

Observed			Predicted		
			Result_of_Treatment		Percentage Correct
			0	1	
Step 1	Result_of_Treatment	0	9	10	47.4
		1	0	71	100.0
	Overall Percentage				88.9

a. The cut value is .500

The accuracy level of the LR on RF determined variables shows a superior result, for predicting the treatment, compared to all approaches discussed so far.

5.6. Results and Discussions

The analysis of disease data set highlights the cause of the disease. The common factors across the different approaches include the age of the patient, assuming all of them had multiple sex partners and used contraceptives and suffered from sexually transmitted diseases. The results from K-means clustering show that age equal to 24 years with first sexual interaction at the lowest age of 16 caused cervical cancer.

The treatment dataset indicates the treatment attributes once the warts are detected. Here, time is a critical factor. The results from K-means clustering show that most likely time equal to around seven months with an age of 32 years showed signs of cancer. The minimum time of occurrence of cervical cancer stands out as three months post detection of warts. There are four cases where cancer got detected after six months. Thus, the time range of 3 to 7 months is most crucial.

The accuracy of the prediction of the disease appeared to be highest with the approach, namely Decision-Tree. The probability of type 1 and 2 error was 12.7% and 3.2% respectively. The association rule associated with thisapproach is:

Schiller's test result is the significant factor for detection of cervical cancer.
For the Schiller's test result = 1, 64.9 % of women have cervical cancer.
For the Schiller's test result = 0, next best predictor is age
If age > 19 and <= 21 then, 6 % of women have cervical cancer
For the age <= 19, next best predictor is STDs : pelvic inflammatory disease
For STDs : pelvic = 0.0, 0.0 % women are attacked with cervical cancer
For STDs : pelvic = missing, 3.1% of women have cervical cancer
For the age > 21, next best predictor is First sexual intercourse
For First sexual intercourse <= 14, 2.6% of women have cervical cancer
For First sexual intercourse > 14, 0.0% of women have cervical cancer

The accuracy of the prediction of the treatment of the disease appeared to be highest with approaches, namely Logistics-Regression and Decision Tree. The Decision-Tree approach yielded 7% type 1 error. The association rule associated with these two approaches are:

For the time <= 5.250, 84.6 % of women get treated.
If time > 5.250 and <= 8.000 then, 100% of women get treated
If time > 8.000 and <= 10.500 then, 83.3% of women get treated
For the time > 10.500, 29.4 % of women get treated.

Decision Tree analysis showed consistency in performance, while Random-forest performed better in a data set with distinct dichotomy. In the case of disease data, the ratio of biopsy=1 and biopsy=0 was 94:6. Hence, the results from RF-based Decision-Tree were not significant. Whereas in treatment dataset, the ratio was 80:20, the results reflected higher prediction accuracy. The accuracy level of the LR on RF determined variables showed a superior result (96.2%) among all methods, for predicting the treatment.

Hence, the alternate hypothesis – "integration of data mining approaches leads to better prediction" stands accepted for a dataset with dichotomy at the ratio of 80: 20 or better. The significant variables when the disease is detected includes – Age, schiller – STD parameters. The critical factor for successful treatment is Time. The critical values of these factors are- Age: 19 years and above; Time: 3 to 7 months; Schiller Test = 1; STD test (of different parameters) = 1.

Appendix I

Table 5.3. Attribute Information of Second Dataset (With 25 Clusters)

Final Cluster Centers																										
	Cluster																									
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Age	41	36	24	35	35	49	30	31	30	28	29	18	29	28	26	36	28	38	19	27	20	36	33	33	42
Number of sexual partners	3	1	1	3	3	2	5	2	3	4	3	7	3	3	3	3	3	2	2	4	2	1	3	4	3
First sexual intercourse	17	22	20	20	17	15	16	20	17	14	19	16	15	16	14	19	15	20	15	17	18	28	16	17	19
Num of pregnancies	4	4	1	2	6	6	4	2	3	4	2	1	3	3	3	3	6	2	2	3	1	1	4	0	3
Smokes	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	1	1	1	0	0
Smokes (years)	0	16	0	0	13	0	0	9	0	0	0	5	0	12	7	0	14	0	0	0	13	16	14	0	0
Smokes (packs/year)	0	5	0	0	3	0	0	5	0	0	0	5	0	6	1	0	2	0	0	0	7	2	1	0	0
Hormonal Contraceptives	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	1	1
Hormonal Contraceptives (years)	10	0	4	0	7	2	0	6	11	0	3	2	0	7	2	7	7	0	0	1	0	0	1	1	3
IUD	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1
IUD (years)	0	0	0	10	0	0	7	0	1	0	4	0	0	0	0	0	4	0	0	0	0	0	0	3	3
STDs	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
STDs (number)	1	3	2	2	1	1	1	2	2	1	2	2	1	1	3	2	2	1	2	2	2	3	1	1	2
STDs: condylomatosis	0	1	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1
STDs: cervicalcondylomatosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs: vaginalcondylomatosis	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
STDs: vulvo-perinealcondylomatosis	0	1	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1
STDs: syphilis	1	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
STDs: pelvic inflammatory disease	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
STDs: genital herpes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs: molluscumcontagiosum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

STDs: AIDS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs: HIV	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0
STDs: Hepatitis B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
STDs: HPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
STDs: Number of diagnosis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
STDs: Time since first diagnosis	21	7	2	3	12	3	11	16	8	8	5	3	16	2	5	18	10	3	2	3	21	1	16	11	20		
STDs: Time since last diagnosis	21	7	2	3	12	3	11	16	8	8	5	3	16	2	5	18	10	3	2	3	21	1	16	11	20		
Dx: Cancer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0		
Dx: CIN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Dx: HPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0		
Dx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0		
Hinselmann	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Schiller	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0		
Citology	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0		
Biopsy	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0		

Chapter 6

Identifying Association Rules for the Assessment of the Possibilities of Cardio Vascular Disease

Introduction

The use of Data Mining Techniques (DMT) in cardiovascular disease studies has shown that DTs, K-means and LR classifiers have obtained acceptable results. This thesis proposed association rules based on any of these single approaches that produce better results than others. Issues remain with any of these techniques and may if used in isolation, affect decisions. The use of any of the techniques may result in Type I error. The approach of this study assesses the ability of three Data Mining (DM) algorithms e.g., K-Means Algorithm, Decision Tree (DT), and Logistic Regression (LR) to identify these key medical tests and to establish association rules between the outcome of these tests and the potential on detection of CVDs.

Methods

The Hungarian Institute of Cardiology is the source of the data used in this study. The CVD data set is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. A dataset of 155 patients who went for medical examinations was randomly selected for subjects. The biometric data collected during the physical examination of the following factors are shown in Table 6.1..

Table 6.1. Description of Dataset

	Attributes	Description	Values
1	AGE	Age in years	Continuous
2	SEX	Male or Female	1=Male, 0 = Female
3	CP	Chest Pain Type	1=typical type, 2=typical type angina, 3=non-angina pain, 4 = aymptomatic
4	TRESTBPS	Rest blood Pressure	Continuous value in mm hg
5	CHOL	Serum Cholesterol	Continuous value in mm/dl

6	FBS	Fasting Blood Sugar > 120 mg/dl	1=true, 0=false
7	RESTECG	Resting Electrocardiographic Results	0=normal, 1=having ST-T wave abnormality, 2=showing probable or definite left ventricular hypertrophy
8	THALACH	Maximum Heart Rate Achieved	Continuous
9	EXANG	Execersise Induced Angina	1=yes, 0=false
10	OLDPEAK	ST Depression induced by exercise relative to rest	Continuous
11	SLOPE	Slope of the peak exercise ST segment	1=unsloping, 2=flat, 3=downsloping
12	CA	Number of major vessels colored by Fluorosopy	0 - 3
13	THAL	Defect Type	3=normal, 6=fixed defect, 7=reversable defect
14	Class	Detection of HD	0 = no heart disease (<50% narrowing)
			1 = has heart disease (>50% narrowing)

Compared to other techniques in literature, DT classifiers, K-Means and LR have achieved acceptable outcomes. In developing patient outcomes, minimizing the cost of medicine, and further advanced clinical trials, this type of analysis can play a vital role. As revealed from the literature review, the above three methods were used to extract the entire set of association rules preventing type 1 error.

Results and Discussions

The simulation results (using Python programming language and SPSS) show these model's potential with good accuracy and subsequent convergence for CVD prediction.

Logistic Regression

Table 6.2. The Output Window Provided a Way of Evaluating Which Variables May Have Predictive Value

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.5683	0.973	-5.721	0.000	-7.476	-3.661
cp	0.6219	0.233	2.668	0.008	0.165	1.079
restecg	0.4466	0.219	2.035	0.042	0.017	0.877
ca	0.9547	0.269	3.547	0.000	0.427	1.482
thal	0.4649	0.112	4.142	0.000	0.245	0.685

The variables in the Table 6.2. include coefficients for the line (fitted) and other relative coefficients data. The line formula derived from the output is shown in equation 6.1 below

$$\log\left(\frac{p}{1-p}\right) = -5.5683 + 0.6219 \times \text{cp} + 0.4466 \times \text{restecg} + 0.9547 \times \text{ca} + 0.4649 \times \text{thal} \quad \cdots (6.1)$$

This figure (Figure 6.1.) indicates that the regression model predicts the dependent variable significantly well. This indicates the statistical significance of the regression model that was run. Here, the significant values of cp, restecg, ca and thal is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

Model Validation

The output result from LR summarizes the accuracy of prediction and other related parameters for group classification are shown below.

$$\text{The accuracy of the model} = \frac{TP+TN}{TP+TN+FP+FN} = 0.74$$

$$\text{The Misclassification} = 1 - \text{Accuracy} = 0.26$$

$$\text{Sensitivity or True Positive Rate} = \frac{TP}{TP+FN} = 0.63$$

$$\text{Specificity or True Negative Rate} = \frac{TN}{TN+FP} = 0.87$$

$$\text{Positive Predictive value} = \frac{TP}{TP+FP} = 0.83$$

$$\text{Negative predictive Value} = \frac{TN}{TN+FN} = 0.68$$

$$\text{Positive Likelihood Ratio} = \frac{\text{Sensitivity}}{1-\text{Specificity}} = 4.69$$

$$\text{Negative likelihood Ratio} = \frac{1-\text{Sensitivity}}{\text{Specificity}} = 0.43$$

Here TP, TN, FP, FN stands for True Positive, True Negative, False Positive and False Negative, respectively.

Decision Trees

Evaluating the Model

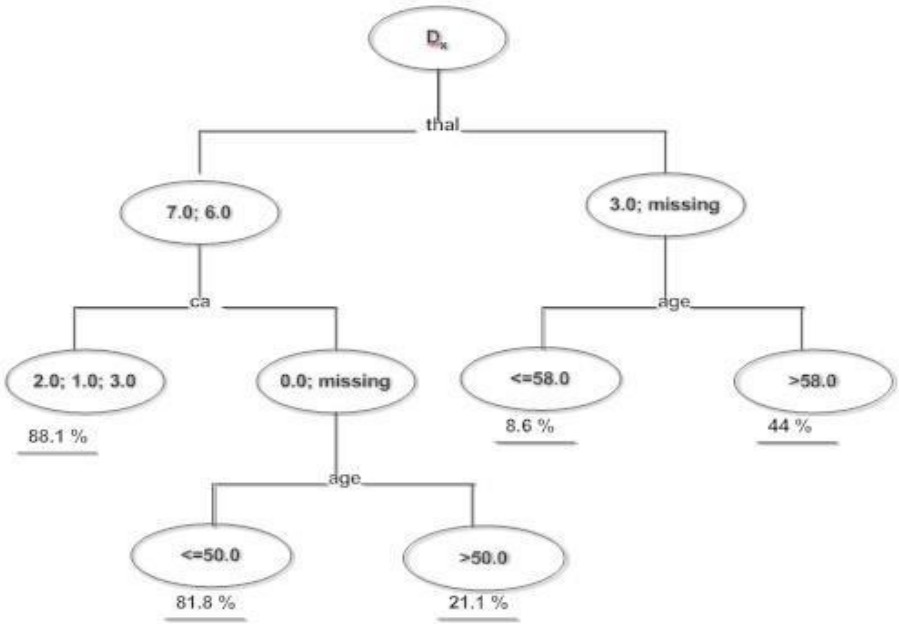


Figure 6.1. DT by CHAID Algorithm

This tree diagram (Figure 6.1.) shows that:

thal factor is the best CVD indicator using the CHAID test.
If value of thal is 6 and 7 the next best predictor for thal is ca and the next best predictor for thal value 3 is age.
For ca values 1, 2 or 3 thal and ca itself are the primary factor of CVDs.
This is known as a terminal node because there are no child nodes below it.
For ca value 0 or missing the next best predictor is age.
If the value of age group is either less than or equal to 50 along with thal value of 7 or 6
then 81.8% of patients can be treated as CVD patients.
If the age group of patients are greater than 50
then 21.1 percent are attacked in CVDs.
On the other hand, if thal's minimum value is 3 or missing then the next best predictor will be age.
For an age group of less than or equal to 58, about 8.6 percent of patients are attacked with CVD
and
For an age level of more than 58, about 44 percent are treated as heart patient.

The performance window (as shown in Table 6.3.) offers a correlation between the outcome observed and the outcome expected.

Table 6.3. Classification Table of DT Method

Observed	**Predicted**		
	0	**1**	**Percent Correct**
0	82	7	92.10%
1	20	46	69.70%
Overall Percentage	65.80%	34.20%	82.60%
Growing Method: CHAID			
Dependent Variable: heartdisease :: category 0/1			

K-Means Algorithm

Analysis with age, cp, ca and thal factor is given.

Table 6.4. Final Cluster Centres

	Cluster				
	1	**2**	**3**	**4**	**5**
age	66	37	43	59	52
cp	3	3	3	3	3
ca	1	0	0	1	1
thal	5	4	4	5	5
Heartdisease :: category 0/1	0	0	0	1	0

The analysis with age, cp, ca and thal factors shows (Table 6.4.) that in the age group of 59 years may suffer from CVD due to cp value 3, and ca is 1 and thal is 5.

Association Rules

Decision Trees

The set of association rule derived from the DT analysis is compiled below

If (6.0<thal<7.0) and (1.0<ca<3.0), Then the risk of CVD is 88%.
If (6.0<thal<7.0) and (ca is missing) and age<=50, Then there is an 81.8 % chance of CVD.
If (6.0<thal<7.0) and (ca is missing) and age>50, Then there is a 21.1% chance of CVD.
If (thal=3.0) and age<=58, Then there is an 8.6% chance of CVD.
If (thal=3.0) and age>58, Then there is a 44% chance of CVD.

Logistic Regression

The set of association rule derived from the LR analysis is compiled below (Equation 6.2).

$$\log\left(\frac{p}{1-p}\right) = -5.5683 + 0.6219 \times \text{cp} + 0.4466 \times \text{restecg} + 0.9547 \times \text{ca} + 0.4649 \times \text{thal} \quad \ldots\ldots\ldots (6.2)$$

The Variables in the Equation table have several important elements. The statistics and associated probabilities provide an index of the significance of each predictor in the equation. The simplest way to assess statistic from the output is to take the significant values and if it is less than 0.05 reject the null hypothesis as the variable does make a significant contribution. In this case, we note that cp(p=0.008), restecg(p=0.042), ca(p=0.000) and thal(p=0.000) contributed significantly to the forecast of CVD but other variables did not (as p>=0.05)).

So, we will drop independents from the model when their effect is not significant by the statistic.

K-Means Algorithm

The set of association rule derived from the K-Means Algorithm analysis is compiled below

If (Age = 59 and ca = 1 and thal = 5 and cp = 3)
Then Probability of CVD is High
else
Probability of CVD is Low

The summary of the results and discussions are presented in the Table 6.5..

Table 6.5. Comparison of Accuracy Level and Identification of Significant Variables (CVD Dataset)

Comparison of Accuracy Level and Identification of Significant Variables for CVD Dataset															
Approaches	**age**	**sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**Accuracy of predicting 1**	**Accuracy of predicting 0**
DT	1										1	1		65.80%	34.20%
LR			1				1				1	1		74.19%	25.81%
K Means	1		1								1	1			
Total	2		2				1				3	3			

Conclusions

The objective of this application is to identify the correct age for individuals suffering from CVDs. A literature review shows that there are several DMTs of varying degrees of accuracy. The above study shows that four variables typically influence CVD (namely age, cp, ca, and thal). Based on these four variables, the insurance companies fix the premiums. Based on the different values of these four variables, the premium slabs will differ according to the probability of CVD. To derive the complete set of association rules leading to the identification of CVDs, the techniques, namely DT classifiers, K-Means, and LR were suggested. The use of any of these methods will likely contribute to the Type 1 error and none will have a 100% accuracy level.

DT analysis shows that factors, namely thal, ca and age are significant to predict the probability of CVD. K-Means classifiers further show the age group for CVD can be lower if a partial variable set is considered. The LR model indicates the probability of CVD as the sum of the product of these variables, namely cp, restecg, ca, and thal with their coefficients 0.6219, 0.4466, 0.9547, and .4649 respectively. That is, LR establishes the weightage of the significant factors. However, a cluster of variables with specific values defines the possibility of CVD. Therefore, the significant factors need to be identified, their weights determined and clustering with age is determined for customizations of insurance premiums. Therefore, the association rule can be established accordingly. Such findings can be the basis for customizing insurance premiums for both the categories of life and health instead of the traditional system of calculating premiums on age-wise slabs.

The use of the K-means Cluster approach indicates that CVD can also be affected by the consideration of partial variable set revealed by people less than 40 years of age. This is true when the interactive effect is taken into account instead of all parameters of three variables, namely sex, ca, and trestbps. Besides, the DT lays down the association rule that reinforces the K-Means analysis findings. The DT results show that the presence of all three modifiable risk factors substantially increases the likelihood of CVD.

However, since the K-Means analysis findings are more precise when age is focused, its

findings are included in the association rule. The LR analysis shows that variables with definite weights such as cp, restecg, ca and thal can lead to CVD. Therefore, it can be concluded that the combination of the most reliable techniques as listed in this study may be used to define the association rule instead of using any one technique.

Chapter 7

Integrated Data Mining Approach Based on the Identification of Important and Contradicting Variables for Analysis of Recurrence of Breast Cancer

Introduction

This study attempted to integrate the DMT based on the significance of variables (or in words, variable importance). The three widely used approaches, namely, Decision-Tree, Logistics-Regression, and Discriminant-Analysis, have been experimented with a different choice of variables. The Logistics-Regression and Discriminant-Analysis have been carried out, separately, considering the entire variable set and with the significant variables identified through Random Forest and Decision Tree Analysis. This application has correlated the results from these analyses with the findings of the integrated investigations. In this chapter, the author tested the proposed framework on databases relating to critical diseases such as breast cancer. Thus, this chapter not only contributes to finding a way to enhance prediction levels using DMT but also tries to contribute to society at large.

In this chapter, an integrated approach combining unsupervised and supervised data mining techniques has been proposed to avoid underestimation of the prevalence of the disease.

Choice of Models

This chapter proposes to use data mining models. Namely, Random Forest, K means clustering, Decision trees, Logistics Regression, and Discriminant Analysis for analysis of the cause of the disease and prediction of the disease.

Data Set Description

In this application dataset on breast cancer available in the UCI database have been used. The data file posses 286 records with ten attributes per record. The lists of attributes are depicted in Table 7.1..

Table 7.1. List of Attributes of Breast Cancer Dataset

Sl. No.	Attributes	Values
1	Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
2	Menopause	lt40, ge40, premeno
3	tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59
4	inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
5	node-caps	yes, no
6	deg-malig	1, 2, 3
7	Breast	left, right
8	breast-quad	left-up, left-low, right-up, right-low, central
9	Irradiation	yes, no
10	Class	no-recurrence-events, recurrence-events

Variable Clustering and Importance

The author used the Random Forest technique for identifying the clusters, importance of variables, and contradicting variables. The original data set were subjected to K-means clustering, Decision-tree (DT) analysis, Logistics-regression (LR), and Discriminant Analysis (DA). These techniques were then used to predict the outcomes of the methods over the revised data set. Two sets of revised datasets were derived. One set comprised variables found relatively significant from RF analysis, and the other obtained after removal of contradicting variables. The authors compared the results to observe the change in the accuracy of prediction. The application performed LR and DA on data set comprising the variables found significant from DT analysis, as well to mark the difference in outcomes.

Random Forest (RF) Analysis

Random forest analysis of the data on disease dataset shows that the tumor-size, breast-quad, deg-malig, breast, and age are relatively important variables (as shown in figure 7.1.). The variables, namely inv_nodes and menopause, contradict the prediction.

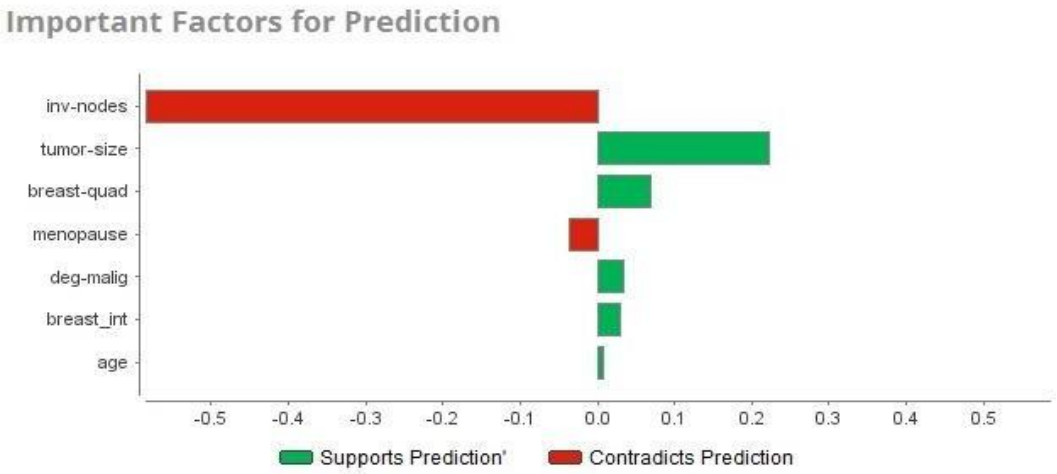


Figure. 7.1. RF Analysis of the Data

K – Means Clustering

Analysis of the Disease Dataset: K – means analysis of the original cancer data set (Table 7.2. and Table 7.3.) show that significant variables include age, tumor-size, node- caps, and breast-quad.

Table 7.2. Final Cluster Centers

	Cluster														
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
age	75	55	58	45	75	59	45	59	45	55	40	65	41	45	34
tumor-size	2	2	30	37	42	19	30	47	46	34	2	22	21	17	37
inv-nodes	1	1	1	16	1	1	1	1	1	16	1	25	1	16	1
node-caps	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0
deg-malig	1	1	2	3	1	2	2	2	2	3	2	3	2	3	2
breast	1	1	1	1	0	0	0	0	1	0	0	1	1	1	1
breast-quad	3	4	2	2	3	2	2	2	1	2	5	1	2	1	2
irradiat	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
Class	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1

The cluster 15 has the highest number of cases (10), which includes the lowest age category of 33.5 years, with comparatively larger tumor size (37). K-means clustering of data set comprising the variables found significant from RF analysis shows that all these variables barring breast are found significant as well (depicted in Table 7.4., Table 7.5. and Table 7.6.).

Table 7.3. ANOVA

	Cluster		**Error**		**F**	**Sig.**
	Mean Square	**df**	**Mean Square**	**df**		
age	1105.577	14	18.85	175	58.651	0
tumor-size	1246.135	14	11.982	175	103.999	0
inv-nodes	132.686	14	0	175	.	.
node-caps	0.338	14	0.052	175	6.516	0
deg-malig	0.892	14	0.529	175	1.688	0.062
breast	0.199	14	0.254	175	0.783	0.687
breast-quad	3.042	14	1.573	175	1.934	0.026
irradiat	0.198	14	0.125	175	1.59	0.086
Class	0.179	14	0.196	175	0.911	0.548

Table 7.4. K-Means on Prominent Variables from RF (Final Cluster Centers)

	Cluster														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
tumor-size	2	2	30	23	42	19	35	52	45	**47**	2	21	22	31	**36**
breast-quad	3	4	2	2	3	2	2	3	1	**2**	5	2	2	2	**2**
deg-malig	1	1	2	2	1	2	2	2	2	**2**	2	2	2	2	**2**
breast	1	1	1	1	0	1	1	0	0	**0**	0	0	1	0	**1**
age	74.5	54.5	64.5	54.5	74.5	65.6	54.5	54.5	44.5	**64.5**	39.5	44.5	34.5	44.5	**33.8**
Class	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	**1**

Table 7.5. ANOVA

	Cluster		Error		F	Sig.
	Mean Square	**df**	**Mean Square**	**df**		
tumor-size	1448.453	14	12.941	239	111.927	.000
breast-quad	3.220	14	1.528	239	2.106	.012
deg-malig	.963	14	.508	239	1.895	.028
breast	.327	14	.246	239	1.329	.191
age	1806.817	14	1.558	239	1159.956	.000
Class	.142	14	.227	239	.625	.843

Table 7.6. Number of Cases in Each Cluster

Cluster	1	1.000
	2	3.000
	3	23.000
	4	45.000
	5	2.000
	6	19.000
	7	35.000
	8	3.000
	9	8.000
	10	**8.000**
	11	4.000
	12	26.000
	13	17.000
	14	45.000
	15	**15.000**
Valid		254.000
Missing		32.000

K-means clustering without contradicting variables (inv_nodes and menopause)

This approach shows that age, tumor size, and breast-quad are significant variables. The variable – node_cap gets eliminated (as shown in Table 7.7., Table 7.8. and Table 7.9.).

Table 7.7. Final Cluster Centers

	Cluster														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
age	74.5	54.5	64.5	60.8	74.5	74.5	54.5	54.5	44.5	64.5	39.5	44.5	34.5	44.5	33.8
tumor-size	2	2	30	18	42	20	28	42	45	47	2	21	22	31	36
node-caps	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deg-malig	1	1	2	2	1	2	2	2	2	2	2	2	2	2	2
breast	1	1	1	0	0	1	1	1	0	0	0	0	1	0	1
breast-quad	3	4	2	2	3	2	2	2	1	2	5	2	2	2	2
irradiat	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
Class	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1

Table 7.8. ANOVA

	Cluster		**Error**		**F**	**Sig.**
	Mean Square	**df**	**Mean Square**	**df**		
age	1774.625	14	3.443	239	515.376	.000
tumor-size	1445.800	14	13.097	239	110.396	.000
node-caps	.149	14	.172	239	.870	.592
deg-malig	.799	14	.518	239	1.542	.097
breast	.258	14	.250	239	1.031	.423
breast-quad	2.864	14	1.549	239	1.849	.033
irradiat	.156	14	.191	239	.814	.653
Class	.171	14	.225	239	.761	.711

Table 7.9. Number of Cases in Each Cluster

	1	1.000
	2	3.000
	3	23.000
Cluster	4	27.000
	5	2.000
	6	2.000
	7	55.000

	8	18.000
	9	8.000
	10	8.000
	11	4.000
	12	26.000
	13	17.000
	14	45.000
	15	15.000
Valic		254.000
Missing		32.000

The largest cluster has 15 cases that are higher than the previous most populated cluster. This cluster includes the lowest age category of 33.5 years, with tumor size 36. The degree of malignancy is two in both cases.

Decision Tree

Figure 7.2. shows the outcome of the Decision tree analysis carried out on the original dataset. The study shows that deg-malig is the primary cause followed by node-caps. For node-caps = 1.0 with deg-malig = 3, 76.7% of patients suffer from recurrence of cancer. The decision tree with all variables predicts the accuracy of 75.9% (depicted in Table 7.10.).

Table 7.10. Accuracy Level of Decision Tree Analysis

Classification			
Observed	**Predicted**		
	0	**1**	**Percent Correct**
0	194	7	96.5%
1	62	23	27.1%
Overall Percentage	89.5%	10.5%	**75.9%**
Growing Method: CHAID **Dependent Variable:** Class (In numeric)			

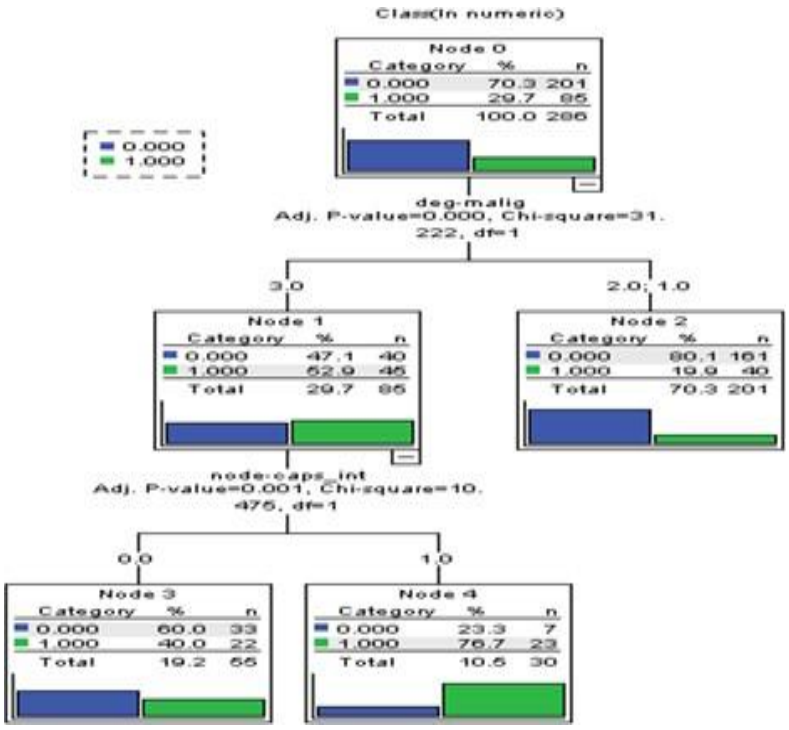


Figure 7.2. Decision Tree for Prediction of Recurrency of Breast Cancer by CHAID

Decision Tree Analysis Considering Relatively Important Variables (Determined from RF Analysis)

Decision tree analysis on variables identified as relatively significant using RF analysis shows that deg-malig is the only critical factor relating to the disease (Figure 7.3.). The overall prediction with this approach reduced to 72% accuracy (Table 7.11.) from 75.9% when predicted with all variables using the decision tree. However, the prediction percentage of recurrence of breast cancer increased from 27.1% to 52.9%, hence type 1 error gets reduced with the integration of results from RF and DT analysis. While the type 2 error increased by 16.5 %. The prediction of a non-recurrence of breast cancer reduced to 80% from the previous prediction of 96.5%.

Decision Tree Analysis after Removal of Contradicting Variables from theOriginal Dataset

The results from this approach show no change compared with the analysis of the original dataset (as shown in Figure 7.3.).

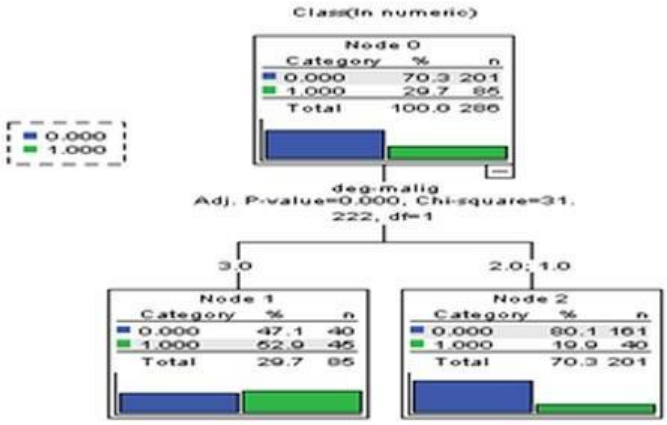


Figure 7.3. Decision Tree Analysis after Removal of Contradicting Variables from the Original Dataset

This tree indicates deg-malig as the only factor for the detection of breast cancer recurrence, but the accuracy of predicting recurrence is higher (52.9%) compared to the earlier DT analysis (depicted in Table 7.11.).

Table 7.11. Accuracy Level of RF-based Decision Tree Analysis

Classification			
Observed	**Predicted**		
	0	**1**	**Percent Correct**
0	161	40	80.1%
1	40	45	52.9%
Overall Percentage	70.3%	29.7%	**72.0%**
Growing Method: CHAID **Dependent Variable:** Class (In numeric)			

Logistic Regression

Logistics Regression on original dataset shows that degmalig and the ninth dummy for tumor size are relatively significant (depicted in Table 7.12.).

Table 7.12. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age			3.149	5	.677			
	age(1)	-20.591	40192.970	.000	1	1.000	.000	.000	.
	age(2)	.084	1.369	.004	1	.951	1.088	.074	15.917
	age(3)	-.398	1.327	.090	1	.764	.671	.050	9.046
	age(4)	-.474	1.276	.138	1	.711	.623	.051	7.596
	age(5)	.192	1.289	.022	1	.881	1.212	.097	15.156
	Menopause			1.573	2	.456			
	menopause(1)	-.487	.484	1.012	1	.315	.615	.238	1.587
	menopause(2)	.414	.956	.187	1	.665	1.512	.232	9.852
	Tumor-size			6.070	10	.809			
	tumorsize(1)	-.728	1.457	.250	1	.617	.483	.028	8.398
	tumorsize(2)	-.746	.911	.670	1	.413	.474	.080	2.827
	tumorsize(3)	-.341	.866	.156	1	.693	.711	.130	3.878
	tumorsize(4)	-.391	.836	.219	1	.640	.676	.131	3.484
	tumorsize(5)	-.330	.838	.155	1	.694	.719	.139	3.714
	tumorsize(6)	-.619	.956	.419	1	.517	.538	.083	3.507
	tumorsize(7)	-.876	.958	.836	1	.361	.417	.064	2.723
	tumorsize(8)	-20.273	19900.004	.000	1	.999	.000	.000	.
	tumorsize(9)	-2.666	1.317	4.097	1	.043	.070	.005	.919
	tumorsize(10)	-.861	1.575	.299	1	.585	.423	.019	9.258
	Invnodes			3.590	6	.732			
	invnodes(1)	-1.223	1.436	.726	1	.394	.294	.018	4.907
	invnodes(2)	-.559	1.655	.114	1	.736	.572	.022	14.656
	invnodes(3)	19.410	40192.969	.000	1	1.000	268880908.808	.000	.

invnodes(4)	-.629	1.467	.184	1	.668	.533	.030	9.454
invnodes(5)	-.280	1.477	.036	1	.850	.756	.042	13.670
invnodes(6)	-.177	1.543	.013	1	.909	.838	.041	17.243
nodecaps	.360	.455	.626	1	.429	1.434	.587	3.501
Degmalig	.655	.232	7.958	1	.005	1.925	1.221	3.035
breast	.354	.331	1.141	1	.285	1.425	.744	2.728
Breastquad			2.881	5	.718			
breastquad(1)	21.477	40192.969	.000	1	1.000	2125504 191.616	.000	.
breastquad(2)	-.930	.811	1.314	1	.252	.395	.080	1.936
breastquad(3)	-.498	.497	1.003	1	.317	.608	.229	1.610
breastquad(4)	-.693	.486	2.034	1	.154	.500	.193	1.296
breastquad(5)	-.984	.749	1.725	1	.189	.374	.086	1.623
irradiat	.327	.359	.827	1	.363	1.386	.686	2.804
Constant	-.080	2.205	.001	1	.971	.923		

a. Variable(s) entered on step 1: age, menopause, tumorsize, invnodes, nodecaps, degmalig, breast, breastquad, irradiat.

The results showed an overall accuracy of 76.6% (Table 7.13.). The prediction of recurrence was low (41,2%) compared that of non-recurrence (91.5%).

Table 7.13. Classification Table[a]

Observed			Predicted		
			Class (In numeric)		Percentage Correct
			0	1	
Step 1	Class (In numeric)	0	184	17	91.5
		1	50	35	41.2
Overall Percentage					76.6

a. The cut value is .500

Logistic Regression Analysis of Disease Dataset Considering Relatively Important Variables Determined from RF Analysis

In this approach, as well, deg-malig and the ninth dummy for tumor size were found relatively significant; however, showed a decreased accuracy level of 73.8%. The prediction accuracy of the recurrence got further lowered. Thus, the integration of RF and LR did not improve the results. Table 7.14. shows the output of LR analysis.

Table 7.14. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	tumorsize			7.173	10	.709			
	tumorsize(1)	-1.194	1.429	.698	1	.403	.303	.018	4.986
	tumorsize(2)	-.811	.883	.843	1	.358	.444	.079	2.510
	tumorsize(3)	-.430	.827	.270	1	.603	.651	.129	3.288
	tumorsize(4)	-.420	.821	.262	1	.609	.657	.132	3.283
	tumorsize(5)	-.219	.814	.073	1	.788	.803	.163	3.960
	tumorsize(6)	-.556	.932	.355	1	.551	.574	.092	3.565
	tumorsize(7)	-.859	.918	.875	1	.349	.424	.070	2.561
	tumorsize(8)	-20.652	19507.566	.000	1	.999	.000	.000	.
	tumorsize(9)	-2.661	1.278	4.333	1	.037	.070	.006	.856
	tumorsize(10)	-.460	1.493	.095	1	.758	.631	.034	11.775
	Breast-quad			2.996	5	.701			
	breastquad(1)	20.675	40192.969	.000	1	1.000	9526819 50.743	.000	.
	breastquad(2)	-.956	.783	1.491	1	.222	.384	.083	1.784
	breastquad(3)	-.391	.476	.674	1	.412	.677	.266	1.720
	breastquad(4)	-.688	.466	2.175	1	.140	.503	.202	1.254
	breastquad(5)	-.689	.673	1.049	1	.306	.502	.134	1.877
	degmalig	.827	.211	15.316	1	.000	2.287	1.511	3.460
	breast	.240	.311	.593	1	.441	1.271	.690	2.340
	Age			3.961	5	.555			

	age(1)	-20.672	4019 2.970	.000	1	1.000	.000	.000	.
	age(2)	.550	1.278	.185	1	.667	1.734	.142	21.228
	age(3)	-.080	1.253	.004	1	.949	.923	.079	10.764
	age(4)	-.344	1.253	.075	1	.784	.709	.061	8.267
	age(5)	.049	1.273	.001	1	.969	1.050	.087	12.731
	Constant	-1.629	1.518	1.152	1	.283	.196		
a. Variable(s) entered in step 1: tumor-size, breast-quad, deg-malig, breast, age.									

Table 7.15. shows the accuracy level of this integrated approach.

Table 7.15. Classification Table

Observed			Predicted		
			Class (In numeric)		Percentage Correct
			0	1	
Step 1	Class	0	187	14	93.0
		1	61	24	28.2
	Overall Percentage				73.8
a. The cut value is .500					

LR without Contradicting Variable

The Logistics-regression on the revised dataset, i.e., after removal of the contradicting variables, shows improvement in overall prediction (depicted in Table 7.16.). The accuracy of prediction of recurrence by methods I and III are equal, while the forecast of non-recurrence by approaches II and III are equal. Thus, the dataset excluding the contradicting variable appears more effective compared a dataset with important variables alone. Here, contradicting variables influences the accuracy significantly. Therefore, this method enables a reduction in both Type 1 and 2 errors. The significant variables include tumor size and deg-malig.

Table 7.16. Classification Table

Observed			Predicted		
			Class (In numeric)		Percentage Correct
			0	1	
Step 1	Class (In numeric)	0	187	14	93.0
		1	50	35	41.2
	Overall Percentage				77.6

a. The cut value is .500

Logistic Regression Analysis on Relatively Important Variables Determined from DT Analysis

The Logistics regression on variable found significant using DT shows that deg-malig and nodecaps, along with the constant, are statistically significant and predicts the recurrence (as shown in Table 7.17.). Equation 7.1 enables the prediction of the disease with these variables.

Table 7.17. Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
	degmalig	.807	.209	14.927	1	.000	2.240	1.488	3.373
Step 1[a]	nodecaps	1.008	.327	9.510	1	.002	2.739	1.444	5.197
	Constant	-2.830	.477	35.218	1	.000	.059		

a. Variable(s) entered on step 1: degmalig, nodecaps.

$$\log\left(\frac{p}{1-p}\right) = -2.830 + 0.807 \times \text{degmalig} + 1.008 \times \text{nodecaps} \qquad \text{.................... (7.1)}$$

The results showed an overall accuracy of 75.9% (Table 7.18.). But this method had very low predictability of recurrence (27.1%).

Table 7.18. Classification Table

Observed			Predicted		
			Class (In numeric)		Percentage Correct
			0	1	
Step 1	Class (In numeric)	0	194	7	96.5
		1	62	23	27.1
	Overall Percentage				75.9

a. The cut value is .500

Thus, LR on the data set without contradicting variables showed better accuracy compared to other approaches so far.

Discriminant Analysis

The Box's test of equality (as depicted in Table 7.19.) of covariance matrices is found significant. Hence this approach with the original dataset is not considered in this study.

Table 7.19. Test Results

Box's M		19.057
F	Approx.	6.291
	df1	3
	df2	528955.537
	Sig.	.000

DA on Variables Found Significant from RF Analysis

This approach is considered as Box's test of equality of covariance matrices is not found significant. The results from DA on the revised disease data set show that eigenvalue is .073 (<1). Canonical correlation, rc=0.261(<0.35) (Table 7.20.) WilksLambda = 0.932, p-value = <0.001(Table 7.21.). Thus, Function 1 explain the variation well with the low canonical correlation between the two canonical variables.

Table 7.20. Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.073[a]	100.0	100.0	.261

a. First 1 canonical discriminant functions were used in the analysis.

Table 7.21. Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.932	17.759	1	.000

The centroid for the discriminant function – F calculated as ((-0.189 + 0.383)/2) = 0.097 is the discriminating value.

Table 7.22. Canonical Discriminant Function Coefficients

	Function
	1
deg-malig	1.416
(Constant)	-2.988

Unstandardized coefficients

Only one variable found significant (Table 7.22.). Equation 7.2 provides the F function.

$$F = -2.988 + 1.416 \times \text{deg-malig} \quad \text{.....(7.2)}$$

Table 7.23. gives the accuracy of the prediction.

Table 7.23. Classification Table

		Class (In numeric)	Predicted Group Membership		Total
			0	**1**	
Original	Count	0	161	40	201
		1	40	45	85
	%	0	80.1	19.9	100.0
		1	47.1	52.9	100.0
Cross-validated[b]	Count	0	161	40	201
		1	40	45	85
	%	0	80.1	19.9	100.0
		1	47.1	52.9	100.0

a. 72.0% of original grouped cases correctly classified.

b. 72.0% of cross-validated grouped cases correctly classified.

DA Without Contradicting Variables

The Box's test of equality of covariance matrices (Table 7.24.) shows that the test result is significant, and hence, the authors do not consider this approach for further study.

Table 7.24. Test Results

Box's M		11.790
F	Approx.	3.890
	df1	3
	df2	673467.575
	Sig.	.009

Tests null hypothesis of equal population covariance matrices.

7.5. Results and Discussions

The analysis using unsupervised and supervised classifiers, and prediction techniques with and without important and contradicting variables separately show higher accuracy of

prediction of recurrence using DT analysis on a dataset with the relatively important variables determined using RF. The dataset excluding the contradicting variables when used results remain unchanged. This approach has the lowest type 1 error; however, DT analysis on the original data set exhibits the lowest type 2 error, related to non-recurrence. That is, the overall performance of the DT was superior compared to the other methods. Figure 7.4. shows the comparison of the performance of the different approaches.

This study highlights that inclusion and removal of contradicting variables do improve the accuracy results. For example, the accuracy of prediction, of recurrence as well as non-recurrence, using Logistics Regression improved on the inclusion of important variables and exclusion of contradicting variables.

The study found the degree of malignancy as a factor with the highest score followed by the tumor-size. Equation 7.3 gives the expression of this finding.

$$DT_i \cap LR_i \cap DA_i = \{\text{Degree of Malignancy}\} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7.3)$$

Where, DT_i represents Decision Tree analysis on original and revised dataset; LR_i represents Logistic Regression analysis on original and revised dataset; DA_i represents Discriminant Analysis on original and revised dataset; i = I represents original dataset; II represents revised dataset comprising relatively important variables only; III represents revised dataset comprising data of all variables except the contradicting variables.

Interestingly, the results from K-means clustering show that age is a significant factor, not recommended by any other approaches. This Degree of malignancy features as significant variable across supervised as well as unsupervised methods when dataset comprised important variables (obtained from RF analysis) only. Equation 7.4 gives the expression of this finding.

$$SL_{ij} \cap UL_{ij} = \{\text{Degree of Malignancy}\} \; ; \forall i = II \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7.4)$$

Where, SL_{ij} represents the supervised learning approach j on an i^{th} data set; UL_{ij} represents unsupervised learning approach j on an i^{th} dataset; i = I represents the original dataset; II

represents a revised dataset comprising relatively important variables only; III represents a revised dataset comprising data of all variables except the contradicting variables.

	Comparison of accuracy Levels and Identification of Significant Variables								
APPROACHES	tumor-size (avg)	breast-quad-int	deg-malig	breast_int	age (avg)	node-caps,	DEG-MALIG	tumor-size(9TH DUMMY)	Accuracy Predicting 1
K-MEANS I	1	1			1	1			
K-MEANS II	1	1			1		1		
K-MEANS III	1	1			1				
DT - I			1			1			27
DT-II							1		52.9
DT - III							1		52.9
LR - I							1		41.2
LR - II							1	1	28.2
LR - III	1						1	1	41.2
DA - I							1		
DA - II							1		
DA - III									
TOTAL	4	3	1	0	3	2	8	2	

Figure 7.4. Comparison of Accuracy Level and Identification of Significant Variables

The set of variables found significant from unsupervised and supervised techniques include minimum values of age, node cap, tumor size, breast-quad-int and deg of malignancy and is expressed, as shown in expression 7.5.

$$SL_{ij} \cup UL_{ij} = \{\text{age, node cap, tumor size, breast} - \text{quad and deg of malignancy}\} \quad \ldots\ldots (7.5)$$

The minimum values of age, tumor size, breast-quad-int, and deg of malignancy are 33.8, 20, and 2 respectively, that cause recurrence of breast cancer. Thus, the approach of integration of data mining approaches and use of relatively significant variables, and elimination of contradicting variables lead to better prediction.

7.6. Conclusion

This research application emphasizes that the addition and exclusion of contradicting variables from the dataset increase the effects of accuracy. Thus, the approach of integration of data mining approaches and use of relatively significant variables, and the elimination of contradicting variables lead to better prediction.

Chapter 8

Feature Selection for Estimation of Disease Progression in the Genetic Algorithm Model with Logistic Regression

8.1. Introduction

It is important to pick a parameter or function when constructing a multivariable regression model. The primary aim of selecting a feature is to incorporate clinically important and statistically meaningful features into the model with the exclusion of noise/redundant features (Tolles & Meurer 2016; Kiezun et al. 2009). For this aspect, a variety of methods are widely used, such as purposeful selection, best subset, stepwise regression, and association rules. Neither of these methods is a perfect solution, particularly in the case of a large number of features in the era of Big Data. In typical cases, a univariate screening to track variables is necessary for the proper selection of variables. Such variables then construct a template for regression. While some significant variables are operating together, they can be ignored by a system that does not have statistical significance when evaluated independently. This problem can be solved by the best subset approach since all possible candidate variables are tested (Zhang 2016). Moreover, in a dataset that contains a small number of variables, this method may be useful, but may not be the right option for a huge dataset with a very large number of features. The step-by-step approach is another common method, but at the end of the day, it is a local research process (Paterlini & Minerva 2010). GA is a heuristic algorithm for search that emulates biological and natural selection methods (Lucasius & Kateman 1993). The population is changing by selection, crossover, and mutation in Darwin's theory. Although the fittest individuals can live and reproduce, the weakest are eliminated from the population. GA creates artificial random populations (chromosomes in the terminology of GA) evaluated by the method of mathematical fitness. When picking, replicating, crossing, and mutating, both for discreet functionality have been successfully used to solve optimization problems. These artificially designed chromosomes are most suitable for the mathematical fitness function.

This chapter provides a guide on how the Genetic Algorithm (GA) is used to provide the best classification precision for variables selected by Logistic Regression (LR). The most popular approach towards clinical research is the LR classification model since in most cases (e.g., diagnosis) the dependent variables are categorical. Some forms of generalized linear models and data sets can be used conveniently by this framework.

Medical institutions produce and collect large volumes of health data owing to the exponential growth of the IT infrastructure in health services and the higher dissemination of medical databases. It is essential to explore the right feature, available due to the rapidly expanding corpus of medical data, to improve the efficiency of patient care, and in the process also reduce cost and time of treatment.

Methodology

Classification of LRs is a very popular, common, and efficient technique in the healthcare sector. Several studies on the treatment of various diseases involving LR models have already been published. Some of these models were planned to be used in previously unknown instances. Traditionally, this problem has been resolved by step-by-step variable selection techniques for logistic regression models. This sequential method significantly restricts the number of models to be examined. Another way would be to look at all of the existing models. The number of possible models is 2n, given the number of variables to choose from, making this exhaustive method ineffective for variables other than small numbers.

In this study, the authors implement an LR dependent variable selection method with randomly generated values for GA models and compare it to standard linear variable selection approaches using diabetes, liver disease, and heart disease patient data sets. In this section, we review the basic concepts behind LR and explain how we applied them to choosing variables. This research aimed to identify combinations of features generating best-performing predictive models for early diagnosis and development of diabetes, liver, and heart disease. In this approach LR was used in this study to select one or more sets of diagnostic test results (features) that can predict disease progression with high accuracy and

GA is used to build prediction models.

Figure 8.1. and Figure 8.2. illustrate the algorithm architecture and the mechanism which combined LR and GA to predict disease status. The user-selected features for the LR have been used as the GA input. The GA used LR tests to simplify and identify the best set of features for the different sets.

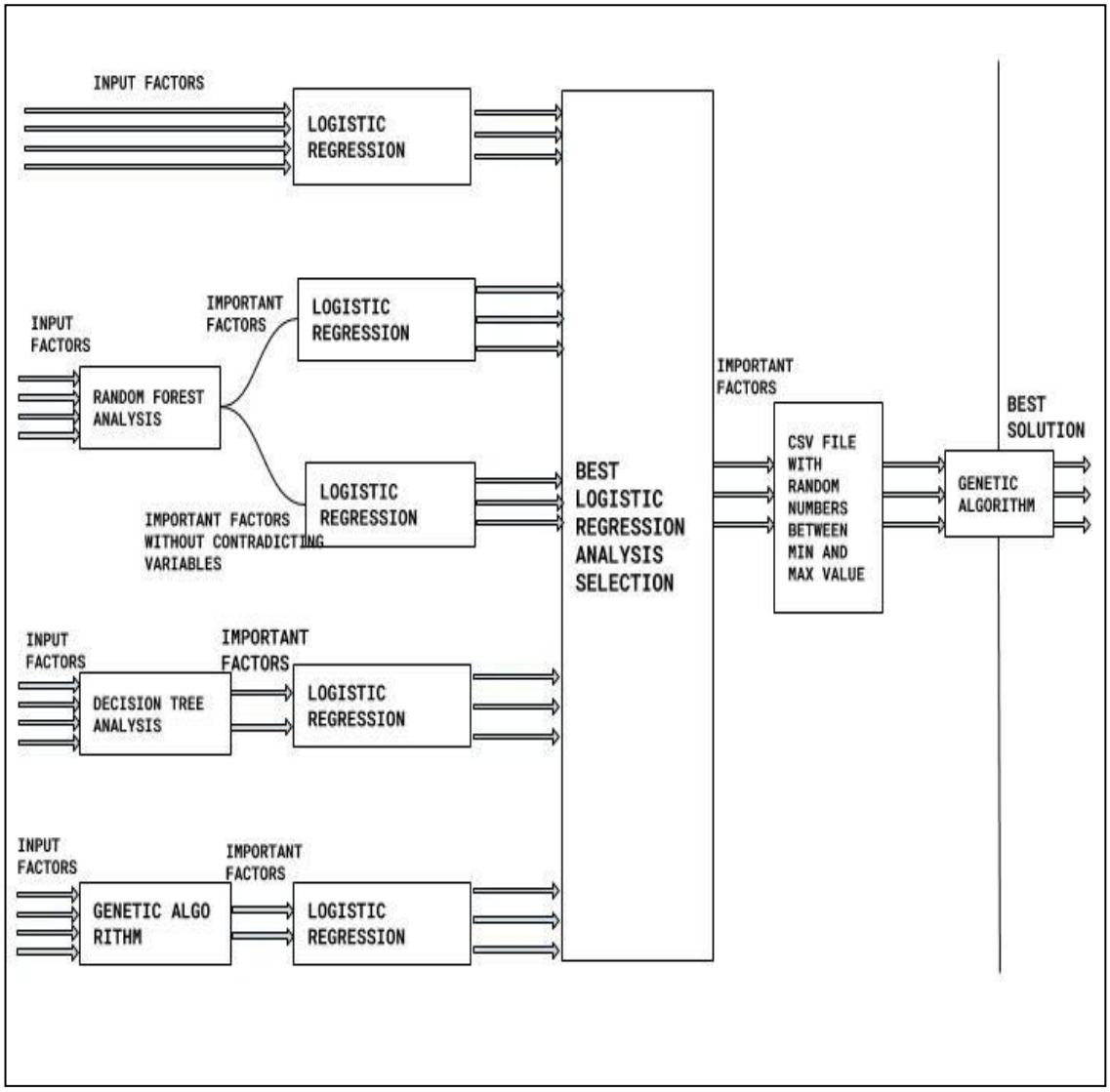


Figure 8.1. The Architecture of the System

8.2.1. The Architecture of the System

In this proposed classifier LR should be paired with GA in the system architecture. The search process included three key steps. Together with the set of selected features from RF (RF-with or without contradictory variables), DT, and GA, LR is used independently to create a predictive model for each instance within possible GA solutions. A new data set will be generated with all predicted variables from the LR's subsequent outputs. Variables between the minimum and maximum values of these variables will be generated randomly. GA search at the last stage would find better solutions that substitute previously found less suitable solutions.

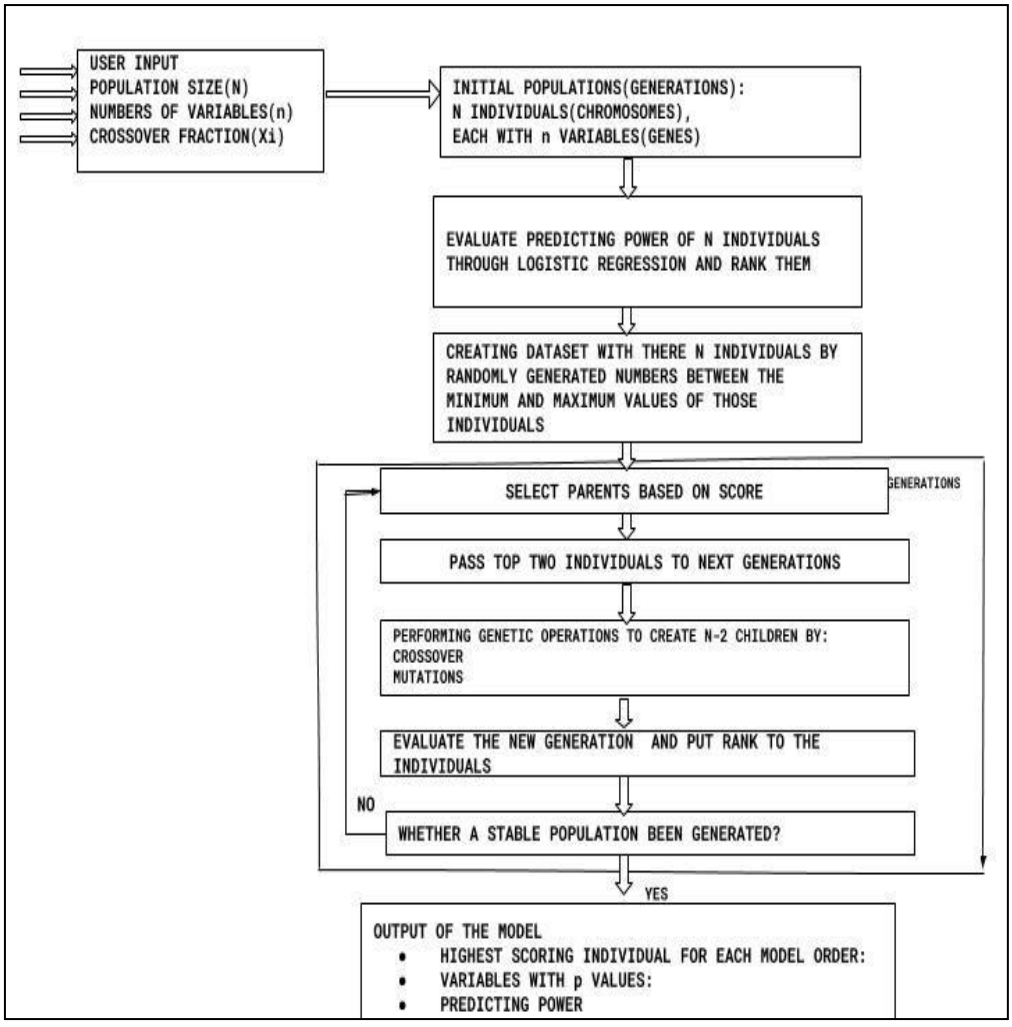


Figure 8.2. A Genetic Algorithm for Disease Outcome Modeling

Attribute Description

The datasets used in this paper are discussed hereafter.

Diabetes Disease Dataset

The author gathers information from the Pima Indians Diabetes Database (www.kaggle.com). This dataset was previously available at the National Institute of Diabetes and Digestive and Kidney Diseases. The basic goal of the dataset is to predict whether or not a patient has diabetes by using different diagnostic procedures, and based on those certain investigative measurements are included in the dataset. To select these instances from a larger dataset, a lot more restrictions were placed. In general, all patients here are Pima Indian heritage females at least 21 years old. The Diabetes data file contains 9 specific dependent and independent attributes (shown in Table 8.1.) collected from patients and a total of 768 records of patients. Out of these 768 patients, 268 have diabetes.

Table 8.1. Description of Diabetes Dataset

	Attributes	Description	Values
1	Pregnancies	Number of times pregnant	Continuous
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	BloodPressure	Diastolic blood pressure	mm Hg
4	SkinThikness	Triceps skin fold thikness	mm
5	Insulin	2-Hour serum insulin	mu U/ml
6	BMI	Body Mass Index	weight in kg/(height in m) ^2
7	DiabetesPedigreeFunction	Diabetes Pedigree Function	
8	Age	Age	Years
9	Outcome	Class variable	0=no diabetes 1=has diabetes

Liver Disease Dataset

In this dataset out of a total of 583 records, 416 records are present for liver-disease patients and 167 persons are non-liver patients. The data are collected from test samples by studying the medical test records of patients from North East of Andhra Pradesh, India, and is available in the UCI repository. Out of the 583 records, 441 are male patients and 142 are female patients. Patients above 89 years are considered 90 years of age. This file contains data about various test samples - their age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos. 'Outcome' is a class label used to divide into groups (liver patient or not). The details of the different features are described in Table 8.2..

Table 8.2. Description of Liver Disease Dataset

Sl. No.	Attributes	Description	Values
1	Age	Age of the patients	Continuous
2	Gender	Sex of patients	1=Male, 0=Female
3	Total_Bilirubin(TB)	Total Bilirubin	mg/dL
4	Direct_Bilirubin	Conjugated Bilirubin	mg/dL
5	Alkaline_Phosphotase(Alkphos)	ALP	U/L
6	sgpt	Enzyme produced by the liver	U/L
7	sgot	Enzyme produced by the liver	U/L
8	Tot_proteins(TP)	Total Proteins	g/dL
9	Albumin(ALB)	Albumin	g/dL
10	Albumin_and_Globulin_Ratio(AGRatio)	A/G Ratio	
11	Outcome	Label (patient has liver disease or not)	0=no lever disease 1=has lever disease

Heart Disease Dataset

Already described in Chapter 7.

Results and Discussions

A GA is defined which searches for the best LR model with only significant variables in the space of all possible subsets of predictor variables. Two out of three datasets with a different number of independent variables showed the approach to be effective with large data set containing hundreds of records and a large number of variables. While the selection of meaningful variables is performed in LR using statistical techniques, as it is a black-box model variable selection cannot be done using GA statistical techniques. In the proposed method, variable selection shall be done by step-by-step LR. Alternatively, explanatory variables derived from LR and projections are taken as GA inputs. Therefore, the proposed method has benefits for both the LR and GA.

When the results are given in Table 8.3., Table 8.4. and Table 8.5. were evaluated, it was seen that in comparison with other methods, the proposed method yielded the best results in terms of criteria for various disease data.

Table 8.3. Result of Analysis for Diabetes Dataset

Method	The accuracy of the model	Selected features
LR	0.79	Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction
LR after DT	0.79	Glucose, BMI, Age
LR after RF	0.8	Glucose, BMI, DiabetesPedigreeFunction , Age
LR after GA	0.79	Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction
Our proposed Model	0.97	Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction

As shown in Table 8.3. we found from different analyses on diabetes data-set that the classification accuracy of the LR after RF is highest, so variables chosen for GA analysis are Glucose, BMI, DiabetesPedigreeFunction and Age. With randomly generated values and LR equation, Y = -8.6545+0.0322 * Glucose + 0.0787 * BMI + 0.83 * DiabetesPedigreeFunction + 0.0302*Age, the prediction accuracy for our proposed model is 97% which outperforms other popular ML classifiers.

Table 8.4. Result of Analysis for Liver Disease Dataset

Method	The accuracy of the model	Selected features
LR	0.75	Age, DB, Sgpt , TP, ALB
LR after DT	0.68	Age, DB, Alkphos
LR after RF	0.67	Age, DB, Sgpt , TP, ALB
LR after GA	0.68	Age, DB, Sgot , TP, ALB
Our proposed Model	1	Age, DB, Sgpt , TP, ALB

As shown in Table 8.4., we found from different analyses on the liver disease dataset that the classification accuracy of the LR is highest, so variables chosen for GA analysis are Age, DB, Sgpt, TP, ALB. With randomly generated values and LR equation, Y = -1.6635+0.0179 * Age + 0.5534 * DB + 0.0157 * Sgpt + 0.4333 * TP – 0.6653 * ALB. The prediction accuracy for our proposed model is 100% which is the highest among all other ML classifiers used in this study.

Table 8.5. Result of Analysis for Heart Disease Dataset

Method	The accuracy of the model	Selected features
LR	0.74	cp, restecg, ca, thal
LR after DT	0.71	ca, thal
LR after RF	0.68	sex, cp, thalach
LR after GA	0.71	thalach, ca, thal
Our proposed Model	0.98	cp, restecg, ca, thal

The results of different analyses on heart disease are shown in Table 8.5. The classification accuracy of the LR is highest, so variables chosen for GA analysis are cp, restecg, ca, thal. With randomly generated values and LR equation, Y = -5.5683 + 0.6219 * cp + 0.4466 * restecg + 0.9547 * ca + 0.4649 * thal. The prediction accuracy with our proposed model is 98% which is the best among all other ML classifiers used in this study.

The proposed model not only generates the highest accuracies in different disease prediction but at the same time also handles the missing value problem in datasets. Generally, medical datasets available in different repositories consist of noise, missing values, and inconsistencies. To overcome these problems, researchers use different pre-processing steps e.g., data cleaning, data integration, data transformation, data reduction, etc. Our new classification model generates numbers between the minimum and maximum values for all factors randomly. This will help to test every possible combination of values among various factors. Additionally, this technique will also overcome the noise, missing value, and inconsistency problems in datasets.

8.4. Conclusion and Future Scope

This research applied GA to predict the progression of the disease by integrating the effects of a large set of independent variables from datasets (n number of possible combinations within the minimum and maximum values) with the best LR formula chosen to calculate the best accuracy. The classification result of this new classifier shows that the combination of variables is superior in quality to the use of a single significant variable or a finite set of variables to predict disease progression. The developed algorithm will also be tested and modified with more data collected to improve the prediction models. The GA algorithm built out of this study was implemented as a general solution that can be applied to other disease or non-disease datasets. GA's approach is designed to identify solutions requiring a successful search of a subset of features to identify almost optimized combinations to solve large, complicated, or poorly understood solution space.

Clinical diagnosis and prognostic outcomes may be considered by choosing the right features

as a matter of classification and reliability. In clinical science, LR has become a more widely used classification technique today. This study demonstrates a way of using GA to estimate accuracies, and choose characteristics in conjunction with LR. Patient survey results, observational retrospective studies, and clinical trials can also help us understand the effectiveness, quality, and cost of this model. Thus, the authors conclude:

1. Feature selection is a necessary condition – to decide whether to keep or discard features,

and

2. Different machine learning techniques (MLTs) can predict with higher accuracy given specific pre-requisites, such as the size of the dataset, number of features, and records per category.

Thus, the author proposed an ensemble of feature selection and machine learning techniques for the prediction of diseases leading to a reduction in errors. However, the attempt to optimize the prediction accuracy by fixing the constraints regarding specificity and sensitivity through a simulation of FS and MLT would yield a better result. The author used such an attempt in next chapters for medical datasets.

Chapter 9

A Novel Enhanced Decision Tree Model for Detecting Chronic Kidney Disease

Introduction

Health issues cannot be put in one basket as some are fully curable, some fatal, and some chronic. Every diagnosis has three distinct features - the accuracy, specificity, and sensitivity.

Electronic health records are now available (e.g., UCI learning repository) and are data intensive. Machine learning derives benefits from the intersection of computer science and statistics, enabling inference drawn from large data sets and present in different forms (Mitchell 2006).

This paper deals with chronic kidney disease (CKD), where data types are varied. They are in ratio scale (e.g., age in years), categorical (e.g., Red Blood and pus Cells: nominal or abnormal; Hypertension: yes or no), integers (Albumin), and non-integers (Specific Gravity). Conventional statistical approaches best suited for similar data types have not been found suitable in the present problem of analyzing CKD. Machine learning enables enhanced learning from the dataset from feature selection, train-test partitioning, the 10-fold cross-validations, and learning algorithms.

Data mining techniques are popularly used in the diagnosis, and no single technique provides consistent results. Results are affected by the dataset's size, selection of the right features, and fitting accuracy. The first criteria are about the data set's availability, and one cannot do anything about it except when simulated to increase the size. The problem aggravates when the line of treatment varies between non-chronic and chronic diseases (Chen et al. 2016; Meza-Palacios et al. 2017).

This research suggests an algorithm that segregates the non-chronic and chronic kidney disease, lists the crucial features causing diseases such as chronic kidney disease (CKD), and improves the accuracy, specificity, and sensitivity of the diagnosis results. Literature reveals very little work regarding the diagnosis of CKD.

This paper answers three primary questions – i. Do the characteristics of patients with kidney disease differ from those suffering from CKD?; ii. Which features are significant for CKD?; and iii. Which method improves diagnosis accuracy and exhibits higher specificity and sensitivity, minimizing Type-1 and Type-2 errors?.

Persistent illness for a long time is a chronic disease. By the definition of the U.S. National Center for Health Statistics, chronic disease has a minimum period of three months or longer. Nowadays, mortality and morbidity are significant factors for chronic disease. In low and middle income-oriented countries, 4 out of 5 deaths are from chronic disease. In India, chronic disease deaths reported in 1990 amounted to 3.78 million (40.4% of all deaths). This figure is expected to grow to 7.63 million in 2020, 66.7% of all deaths (World Health Organization 2011). Chronic Kidney Disease (CKD) is one kind of chronic disease denoting recurring reduction of kidney function. This reduction in kidney function will gradually take place throughout the months or years based on the patients' living conditions (Noia et al. 2013; Hasan & Hasan 2019; Besra & Majhi 2019). CKD is sometimes referred to as a chronic renal failure, where, according to the latest medical estimates, 10% of the world's population is affected by CKD (Levey et al. 2007; Jha et al. 2013). Approximately 58 million deaths were reported worldwide in 2005. According to the World Health Organization (WHO), 35 million people were attributed to chronic diseases. One in five men and one in four women between 65 and 74 are currently estimated to be affected by CKD worldwide. According to the 2010 Global Burden of Disease Survey, CKD ranked 27th in the list of causes for the total number of deaths worldwide in 1990 but sadly fell to 18th in 2010. The extent of progress was second only to HIV and AIDS (Jha et al. 2013).

CKD is a generalized term deals with its structure and functions and can be a source of various diseases with impaired renal function and leads to kidney failure in the advanced stage of it (e.g., albuminuria) after three or more months (Stevens & Levey 2009). Kidney failure is the last CKD stage and developed from the continuously reduced renal activity and renal function complications. Dialysis or kidney transplantations are the only treatment for such patients with end-stage renal disease. The reduced GFR will also increase the risk of cardiovascular disease (CVD), acute kidney injury, infection, cognitive impairment, and

physical function impairment (James et al. 2010; Kriplani et al. 2019; Zeynu & Patil 2018). The complications may be there at any stage of CKD and may result in the patient's death without the kidney failure, or the patient may die due to adverse effects caused due to interventions to prevent or treat CKD (Levey & Coresh 2012).

The CKD diagnosis begins with some clinical tests performed in the laboratory, imaging, or biopsy. Although the biopsy is one of the most common procedures, it has several drawbacks, e.g., it is invasive, time-consuming, expensive, and sometimes dangerous. For example, during the biopsy, the patient may encounter infection, operation risk, and misdiagnosis. The imaging procedure is one of the old, traditional, and effective measures to diagnose kidney disorders. This method also has some drawbacks, e.g., in their use, specifically the harmful effects of radiation. Besides being dangerous, imaging data are not adequate to diagnose CKD (Hasan & Hasan 2019).

Researchers have been motivated to diagnose diseases without the intervention of medical practitioners. This research article has represented the analysis of different previous approaches, and the present study used six supervised machine learning techniques, i.e., Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and proposed Enhanced Decision Tree (EDT) with Feature Selection (FS) algorithm. Recursive Feature Elimination (RFE) performed the FS task for the Kidney disease dataset. For the evaluation of the classifier, train-test partitioning and the 10-fold cross-validations are used.

The proposed approach ensembles feature selection with Enhanced Decision Tree (EDT) classifier. EDT is an ensemble machine learning technique, which provides help in the decision-making process by splitting node attributes randomly. EDT has the potential to predict and diagnose disease. It exhibits tradeoffs between bias and variance, reduces overfitting, and yields high accuracy. The model has low complexity (O(n log n)) and not influenced by outliers. Typically, a method suggests removing outliers as it affects the results; however, removing the record is a loss of information. Other methods such as Decision-tree or Random-forest show high and medium variance, compared to EDT, meaning low

instances or over-fitting. EDT has similar nature to that of Random-forest but builds multiple trees based on samples done without replacement. The split of EDT nodes is random, not based on data bootstrapping, making it robust for classification applications.

This paper has used crucial test statistic, namely, Kappa (Chalak et al. 2020), ROC-AUC (Fan et al. 2006), and Wilcoxon rank-sum test (de Barros et al. 2018), to judge the significance of the method's result.

Feature Selection and Data Mining

Feature selection (FS) in health care ensure the choice of the right tests to diagnose the disease, reduce the cost of treatment, and minimize computation complexity. FS's importance is felt in high dimensional space where data gets sparser and affects algorithms suitable for low-dimensional space; FS helps feature reduction, or otherwise, models may tend to overfit and enhance learning performance (Li et al. 2017). FS are broadly categorized as supervised, unsupervised, and semi-supervised techniques. Supervised methods are designed for classification or regression algorithms. It comprises filter methods independent of learning algorithms and wrapper technique that uses a classifier or a regression model to judge the significance of features. Un-supervised methods are suitable for un-labeled data where the evaluation of feature importance is difficult. This approach uses all instances available and rely on learning methods (wrapper methods) or are independent of learning algorithms (filter approach) or embedding FS in unsupervised learning algorithms (embedded methods). The filter approach may not select optimal features for a specific learning algorithm due to the absence of a specific learning algorithm. Wrapper methods suffer for large features as the search space required is twice the number of features (Li et al. 2017). Embedded methods tradeoff between filter and wrapper methods; however, these are now supplemented with hybrid methods (Guyaon and Elisseeff 2003; Kohavi and John 1997; Arai et. al 2016; Narendra and Fukunaga 1977; Golberg 1989).

Di Noia et al. (2013) used an ensemble of artificial neural networks (ANNs) to identify patients' health status, leading to end-stage kidney disease (ESKD). Two new updated

Boolean Particle Swarm Optimization algorithms, namely Velocity Bounded Boolean Particle Swarm Optimization (VbBoPSO) and Improved Velocity Bounded BoPSO (IVbBoPSO), were implemented by Gunasundari et al. (2016) to solve the issues of feature selection. Experiments are conducted to pick elite features from the data on liver and kidney cancer.

Dataset and Attributes

The dataset used in this research was obtained from the UCI machine learning dataset (Rubini 2015). The dataset contains pathological data of 400 people from the Southern part of India. There are 24 features present in the dataset, most of which are physiological and clinical. Table 9.1. lists various parameters and their data types. The missing values of all attributes have been replaced by the arithmetic mean of the numerical and discrete integer values of all instances in the data pre-processing stage.

Table 9.1. Dataset Description of CKD

Sl. No.	Attribute	Descriptions	Values
1	age	Age of the patient when diagnosed (numerical)	Years
2	bp	Blood Pressure (numerical)	mm/Hg
3	sg	Specific Gravity (nominal)	1.005,1.010,1.015,1.020,1.025
4	al	Albumin (nominal)	0,1,2,3,4,5
5	su	Sugar (nominal)	0,1,2,3,4,5
6	rbc	Red Blood Cells (nominal)	normal, abnormal
7	pc	Pus Cell (nominal)	normal, abnormal
8	pcc	Pus Cell clumps (nominal)	present, not present
9	ba	Bacteria (nominal)	present, not present
10	bgr	Blood Glucose Random (numerical)	in mgs/dl
11	bu	Blood Urea (numerical)	in mgs/dl
12	sc	Serum Creatinine (numerical)	in mgs/dl
13	sod	Sodium (numerical)	in mEq/L

14	pot	Potassium (numerical)	in mEq/L
15	hemo	Hemoglobin (numerical)	in gms
16	pcv	Packed Cell Volume (numerical)	nominal
17	wc	White Blood Cell Count (numerical)	in cells/cumm
18	rc	Red Blood Cell Count (numerical)	millions/cmm
19	htn	Hypertension (nominal)	yes, no
20	dm	Diabetes Mellitus (nominal)	yes, no
21	cad	Coronary Artery Disease (nominal)	yes, no
22	appet	Appetite (nominal)	good, poor
23	pe	Pedal Edema (nominal)	yes, no
24	ane	Anemia (nominal)	yes, no
25	class	Class (nominal)	ckd, notckd

The patients' ages vary from 2 to 90 years, with an average value of 51.48 and a standard deviation is 17.17. Most of the 24 attributes displayed in Table 2 were not used by most state-of-the-art algorithms previously for CKD detections, such as age, serum creatinine, albumin, and urea. These characteristics are described below and their relationship to kidney disease.

Serum creatinine is a muscular activity-based waste material, where the activities extract creatinine entirely from the blood while the kidneys function properly. Creatinine levels in the blood increase with reduced kidney activities. Researches (Levey et al. 1999; Levey et al. 2009; Alaiad et al. 2020; Elhoseny et al. 2019; Qin et al. 2019) show that the most commonly used prediction factors for CKD diagnosis are serum creatinine, age, serum urea, and specific gravity. From different research studies performed in different random samples, it has been observed that close associations are present between enhanced albuminuria (presence of albumin in the urine) and kidney failure (Chronic Kidney Disease Prognosis Consortium 2010). Besides, data from the United States' overall population indicate that albuminuria is the most common CKD marker in young adults (Gansevoort et al. 2011).

Kidney functions reduce drastically by the destruction of blood vessels inside it. This destruction is caused due to high blood pressure. High blood pressure spreads blood vessels and eventually weakens them. Once the blood pressure has increased, the vessels spread

faster, so the blood flows faster inside the kidney. The kidney's damaged blood vessels can prevent waste and excess fluid from being filtered out of the body. If the blood vessels have damaged, the blood vessel's additional fluid may increase blood pressure, further leading to a hazardous cycle (Salekin & Stankovic 2016). CKD is a different coronary artery disease risk factor (CAD). It is the leading cause of death and mortality among CKD patients (Cai et al. 2013). Coresh et al. (2001), in their research, found that 70% of people with an elevated level of serum creatinine have high blood pressure. This high blood pressure causes damage to the blood vessels, which will increase the blood pressure further. Elevated blood pressure, CAD, and hypertension are robust diagnostic features for CKD.

Anemia is a health condition whenever very few red blood cells are present in the blood compared to the blood cells' total count. The primary function of red blood cells is to carry oxygen to the body cells and organs, specifically in the brain and heart. With anemia, all the body cells, brain, and heart will face a shortage of dietary energy or oxygen. Persons attacked in CKD have partial loss of kidney function, which results in anemia in the earlier stage. Anemia may be developed at the early stages of a CKD if someone has 20% - 50% of normal kidney function. The risk factor for early kidney failure is anemia. Hemoglobin, red blood cell count, and mixed-blood cell concentration in patients have been identified in anemia's early stage.

Diabetes is another crucial risk factor for CKD. Research studies have shown that around one-third of diabetic patients also suffer from CKD (Salekin & Stankovic 2016). Packed tiny blood vessels inside the kidney perform the filtration function. A high level of sugar in the blood (diabetes) makes those blood vessels narrow and clogged. Without sufficient blood, the kidneys are impaired, so the albumin goes through these tubes and ends where it should not be in the urine. This condition will reduce blood flow inside the kidney and also worsen the functioning of the kidney. Diabetes also causes severe damage to nerves, due to which diabetic patients cannot feel that the bladder is full or empty. The pressure created due to the full bladder recursively damages the kidney and further worsens the kidney function. Albumins in urine, glucose in the blood are some essential markers for CKD.

The two significant predictors of CKD are sodium and potassium. Balancing potassium and sodium are of utmost importance for the human body. The excess level of sodium and potassium in the body increases the fatality rate among CKD patients. A person attacked with CKD will neither be able to remove potassium, sodium, and fluid from the body nor be able to accumulate into the bloodstream and body cells gradually. An elevated sodium level causes high blood pressure (Salekin & Stankovic 2016). Simultaneously, patients with an advanced CKD level suffer from high potassium levels in the bloodstream called hyperkalemia. Hyperkalemia results in numbness, fatigue, nausea, lower limb swelling, foot ulcer, chest pain, less self-confidence, anxiety, or slow pulse rate among CKD patients.

The medical term for swelling is edema. Edema may occur if small blood vessels leak fluid to the neighboring tissues. The additional liquid builds up, causing the tissue to swell. Severe pedal edema is a result of a kidney disease called nephrotic syndrome.

It has been observed that the primary reason for kidney failures is bacterial and virus infections. Tuberculosis infection is of significant concern in India and Middle East countries. Likewise, Africa is suffering from streptococcal bacterial infections and HIV, Hepatitis B, or Hepatitis C virus infections. In Africa and Latin America, schistosomiasis, leishmaniasis in Africa, and Asia, malaria in Africa are also the main factors for CKD. Those infections generate pus cells in the urine, which is one of the main symptoms of kidney infection.

Methodology

This study proposes to select the significant features before its classification using the recursive feature elimination (RFE) technique, followed by the proposed method - enhanced decision Tree (EDT).

This research compares the EDT method and prediction models and classifiers, namely logistics regression (LR), decision tree (DT), random forest (RF), Naive Bayes (NB), and support vector machine (SVM).

Recursive Feature Elimination (RFE)

RFE is a feature subset selection (FSS) technique that selects the most relevant subset of variables in place of individual feature selection. It uses a wrapper algorithm to select optimal feature subjects, as proposed by Kohavi & John (1997). Figure 9.1. illustrates the methodology for feature selection.

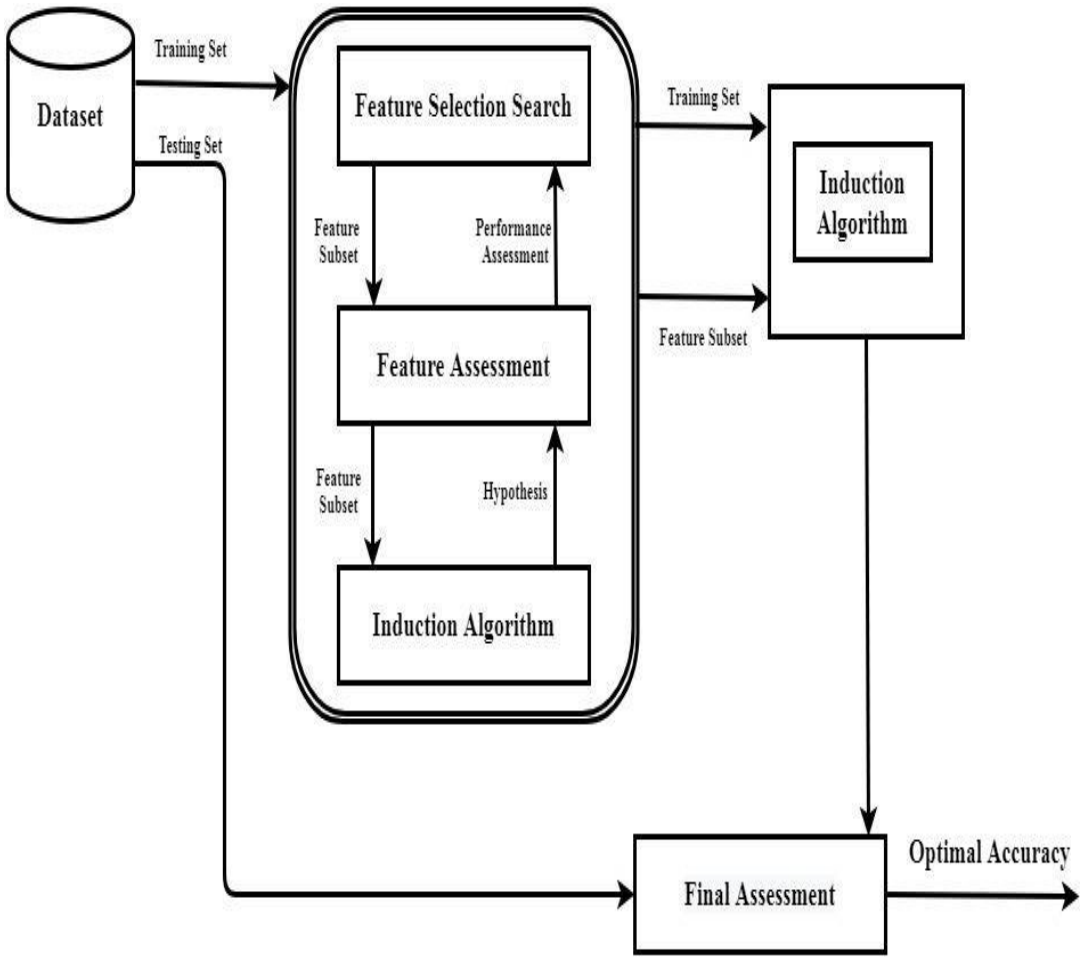


Figure 9.1. Wrapper Algorithm Approach for FSS

The RFE method, proposed here, uses an iterative function (Guyon et al. 2002; Escanilla et al. 2018) that removes the lowest-performing subset of features (using a wrapper) in the process, retains the optimal subset. This subset leads to the maximum accuracy of the output

of the induction algorithm. Thus, in a way, it is a backward greedy algorithm (Kohavi 2000), as illustrated in Figure 9.2., where F^s is set of available features in the data set.

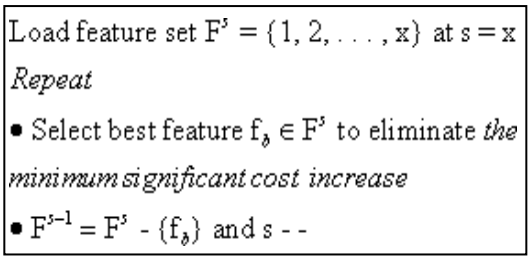


Figure 9.2. Backward Greedy Algorithm (Lee et al. 2012)

f_b denotes the best feature subset selected whose inclusion or removal leads to a change in the estimate's value. The algorithm uses an induction algorithm to search a suitable subset. The subsets are evaluated using n-fold cross-validation (Breiman et al. 1984; Weiss & Kulikowski 1991). The data set is split into n equal (approximately), where an n-1 partitioned set is used as a training set, and the other serves as the test set. The induction algorithm is run n times, and the results are averaged to estimate the accuracy. The induction algorithm should possess the ability to test the outcomes on the validation sets, thus requiring no knowledge of the algorithm.

In the proposed method – backward greedy algorithm, the algorithm starts with all features and removes the feature that most improves or degrades the performance slightly. Most see whether the estimate's value, that is, CKD's prediction accuracy, is reduced. If so, then the feature is removed else marked as f_b.

This approach of feature selection ensures the optimal interaction of the algorithm with the training set, minimizing errors and are straightforward (Draper & Smith 1981; Neter et al. 1990).

Enhanced Decision Tree (EDT) Classifier

The proposed model, Enhanced Decision Tree (EDT) classifier, is an ensemble machine learning technique that randomly split node attributes in the decision-making process. It is an ensemble of decision trees like bootstrap aggregation and random forest technique. The EDT classifier generates many un-pruned decision trees to train the dataset, as depicted in Figure 9.3.. It predicts by calculating the decision trees' average prediction in regression and voting for classification (Landwehr et al. 2005; Breiman 2001; Sharaff and Gupta 2019).

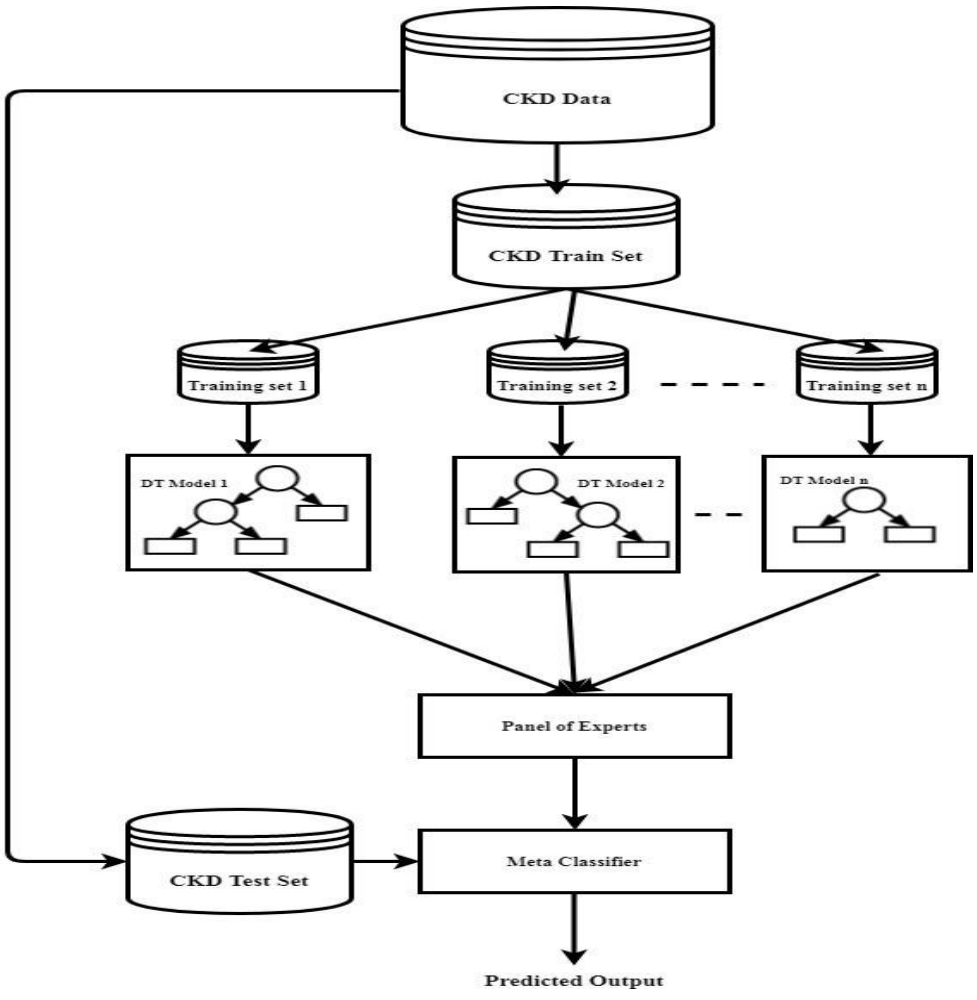


Figure 9.3. Proposed Stacked Model

Method – Description

Figure 9.4. shows the complete architecture of the proposed model. In the first stage of classification, the redundancy present in the dataset's attributes is removed using RFE. This step is followed by prediction using classifiers, namely, LR, NB, SVM, DT, RF, and EDT. The classifiers' outcome with three different train-test split ratios and cross-validation are tested using performance evaluation metrics, namely, confusion matrix, kappa statistics, ROC-AUC, and Wilcoxon analysis.

Performance Metrics

The confusion matrix presents the actual presence of CKD class among the patients and the CKD class's prediction accuracy. This matrix calculates accuracy, specificity, and sensitivity. Accuracy is measured by the ratio of the right (true) predictions with all (true+false) predictions (Zhang et al. 2014). The ratio of True Positive (TP) with False Positive (FP) plus TP is termed as sensitivity (Zhang et al. 2014). Specificity is represented by the ratio of True Negative (TN) with TN+FP. Confusion Matrix is utilized to evaluate the performance of a learning model. Four terms related to the confusion matrix are applied to establish the performance matrices. The number of CKD patients classified as CKD patients are True Positive (TP). False Positive (FP) is the number of non-CKD patients classified as cervical cancer patients, and true negative (TN) is the number of patients classified as non-CKD patients without cervical cancer. False-negative (FN) is the number of patients classified as CKD patients without CKD (Ray & Chaudhuri 2021).

Method – Algorithm

At the outset, the classification is done on the dataset with all features to compare the proposed method's performance. This step serves as the baseline for comparing the performance of the algorithm proposed in this study. Next, the dataset with reduced features (selected) divides dataset into different test-train and ten-fold partitions. This stage avoids the problem of under and overfitting.

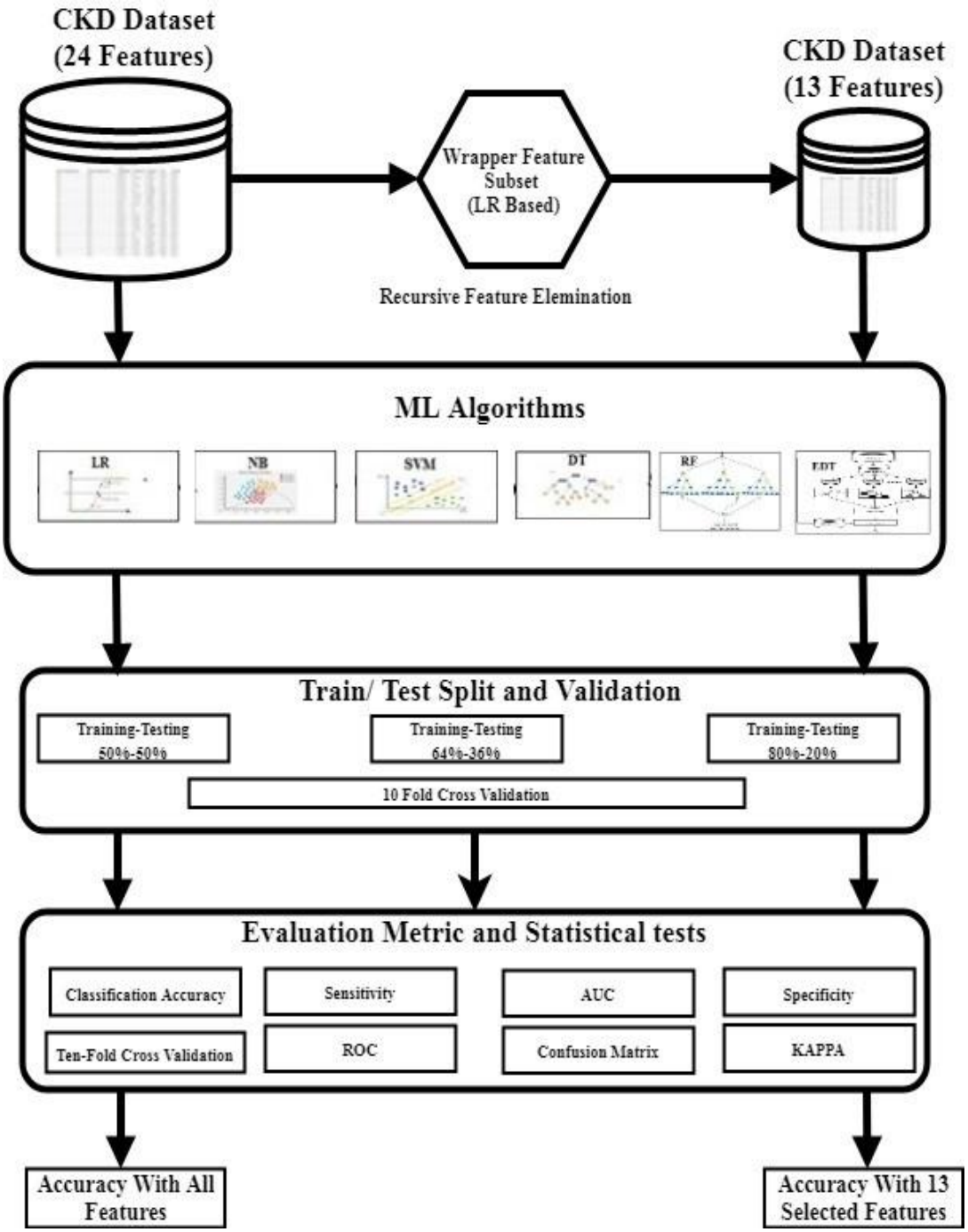


Figure 9.4. Architecture of the Proposed Model

The step-by-step procedure represented in Figure 9.4. is explained below:

Step 1: (Dataset Acquisition) All records from the CKD dataset is collected and read.

Step 2: (Classify CKD and NON-CKD without feature selection) Classification algorithms, namely LR, NB, SVM, DT, RF, and EDT, are used to measure the classification accuracy of the CKD patients.

Step 3: (Selection of relevant features) Backward feature elimination techniques (RFE) is applied to get the relevant feature.

Step 4: (Classify CKD and NON-CKD) Classification of optimal feature subset using algorithms such as LR, NB, SVM, DT, RF, and EDT and accuracy measured.

Step 5: (Validation) The classifiers are trained using the validation set. A test/train partitioning and 10-fold cross-validation technique are used for testing purposes.

Step 6: (Performance parameter computation) Computation of various accuracy parameters - accuracy, sensitivity, specificity, ROC-AUC, Kappa score, and Wilcoxon analysis.

The complexity of the proposed algorithm can be expressed as shown in expression (A).

$$W(n) \in O(n \log n) \quad \text{..... (A)}$$

Where, n is the number of subsets and N denotes a particular subset. The complexity arises out of splitting algorithm as shown in Table 9.2..

Table 9.2. EDT Splitting Algorithm (for Numerical Values of Attributes)

$Split\_node(N)$
Input : Local analysis subset N which is the node that we would like to split
Output : a split $[x < x_p]$ or nothing
If $Stop\_split(N)$ = TRUE then return null.
Else select M attributes $\{x_1, \ldots, x_M\}$ *for* all variable (in N) candidate attributes;
Illustrate M splits $\{n_1, \ldots, n_M\}$, where $n_i = Select\_Random\_Split(N, x_i), \forall i = 1, \ldots, M$;
Return$(splits_*)$ *where* $\text{Score}(n_*, N) = \text{maximum}_{i=1,\ldots,M}\text{Score}(n_i, N)$.

Select_Random_Split(N,x)

Input: subset N and the attribute x

Output: the split

Let $x^{N}_{\max imum}$ and $x^{N}_{\min imum}$ *represents the* $\max imum$ *and* $\min imum$ *values of* x *in* N;
Illustrate a random cut-point x_p homogeneously in [$x^{N}_{\max imum}$ and $x^{N}_{\min imum}$];

Return(split [$x < x_p$]).

End _ Split(*N*)

Input : Local analysis subset N

Output : a true or false value

If $|N| < y_{minimum}$, then return true;

If all attributes are non-variable in N, then return true;

If the output is non-variable in N, then return true;

Else, return false.

The algorithm has two parameters: M, the number of randomly selected attributes for each node, and the minimum sampling size $y_{minimum}$ to separate a node. The (full) original learning sample is used several times to create an ensemble model (The number of trees in this ensemble is denoted by K). In classification problems, tree projections are aggregated by majority vote, and the arithmetic average in regression problems is computed to make the final prediction.

Considering the point of bias-variance, the motivation behind the proposed ensemble classifier is that the explicit randomization of the cut-point and attribute combined with the ensemble average should be able to reduce variance more strongly than the weaker randomization schemes used by other methods. However, given the node splitting procedure's simplicity, the constant factor is much smaller than in other ensemble-based methods that locally optimize cut-points. Considering the point of computational complexity for the balanced trees, the complexity of the growing procedure of decision trees is, on the order of (n log n), i.e., relating to sample size of learning, like most other tree-growing processes. However, given the simplicity of the node splitting procedure, the constant factor is much smaller than in other ensemble-based methods that locally optimize cut-points.

M, $y_{minimum}$ and K parameters have different effects: M determines the intensity of the attribute selection process, the strength of the average output noise is $y_{minimum}$, and K determines the strength of the variance reduction of the ensemble model aggregation. These parameters may be adapted manually or automatically to the problem's details (e.g., by cross-validation).

Like those used in random forests, the EDT algorithm can decrease the variance and bias more strongly than other randomization systems. The variance is created by the model's excessive sensitivity to small fluctuations in the training dataset (high variance can cause overfitting) and is decreased because of the explicit randomization in selecting the features sub-set and the cut-point selection. On the other hand, the full original training dataset is used to learn each decision tree, and the bias, which can be measured as the capacity to correctly generalize unseen data (high bias can cause underfitting), is minimized. Besides, the proposed RFE-based EDT technique does not require high computing time for detecting CKD.

The tuning hyperparameter for the EDT algorithm is the number of decision trees used in the ensemble. The pseudo-code used in the paper includes:

```
edt_model = EnhancedTreeClassifier (n_estimators = num_trees, max_features =
max_features)

edt _model.fit(X_train, y_train.ravel())
```

Typically, the number of trees is increased until the model performance stabilizes. Bagging and EDT algorithms appear to be somewhat immune to overfitting the training dataset given the learning algorithm's stochastic nature.

The number of trees can be set via the "n_estimators" argument and defaults to 100. The set of parameters used in this paper are seed = 7, num_trees = 100, max_features = 13.

Results and Discussions

In this research work, the proposed model is simulated using Python programming language. The training and testing set partitions used are summarized in Table 9.3.. A comparative analysis of six cutting-edge ML algorithms, namely LR, NB, SVM, DT, RF, and EDT, is carried out in this study. Some of these six standard ML techniques improve the accuracy, while others are less efficient. This research work has used the ML ensemble's advanced techniques to boost CKD detection precision and efficiency.

Table 9.3. Training and Testing Set Partition

Training-Testing Partition	Total Training Records	Positive Records in Training Set	Negative Records in Training Set
50-50	200	122(61%)	78(39%)
66-34	264	163(61.7%)	101(38.3%)
80-20	320	198(61.9%)	122(38.1%)
10-fold cross validation	400	250(62.5%)	150(37.5%)

The proposed ensemble ML technique for the EDT classifier for the UCI CKD dataset was 100% accurate. Training and Testing partitions are shown in Table 9.3.. The various methods shown in Table 9.4. and Table 9.5. provided different accuracy levels since the SVM showed 77%, and the EDT with RFE showed 100% accuracy.

Table 9.4. Comparison of Accuracies with All 24 Input Features

Training-Testing Partition	Accuracy with all features					
	LR	NB	SVM	DT	RF	EDT
50-50	0.93	0.96	0.99	0.99	1	1
66-34	0.91	0.96	0.96	0.99	0.99	1
80-20	0.93	0.99	0.98	1	1	1
10-fold cross validation	0.93	0.96	0.63	0.97	1	1

Table 9.5. Comparison of Accuracies with Selected 13 Input Features

Training-Testing Partition	Accuracy with selected features					
	LR	**NB**	**SVM**	**DT**	**RF**	**EDT**
50-50	0.93	0.90	0.98	0.98	0.97	1
66-34	0.99	0.90	0.98	0.98	0.99	0.99
80-20	0.99	0.90	0.98	0.98	0.99	0.99
10-fold cross validation	0.96	0.91	0.77	0.98	1	1

Tables 9.4. and 9.5. provide the accuracy considering all and selected features across all train-test split. The results show the robustness of EDT that performs consistently well and, at the same time, enables the reduction of features, narrowing down the focus of treatment. Table 9.6. and Table 9.7. show each classifier's standard deviations and accuracy, and EDT stands out the best.

Table 9.6. Comparison of Standard Deviation with All 24 Input Features

Training-Testing Partition	Accuracy with all features					
	LR	**NB**	**SVM**	**DT**	**RF**	**EDT**
Standard Deviation (10-fold cross validation)	0.046	0.045	0.46	0.01	0.01	0.0

Table 9.7. Comparison of Standard Deviation with Selected 13 Input Features

Training-Testing Partition	Accuracy with selected features					
	LR	**NB**	**SVM**	**DT**	**RF**	**EDT**
Standard Deviation (10-fold cross validation)	0.02	0.06	0.25	0.02	0.02	0.01

The results of a confusion matrix using different ML algorithms are given in Table 9.8. and Table 9.9.. The performance of EDT based on a subset of features using RFE has the highest values – accuracy = 100%, sensitivity = 1, specificity = 1.

The LR's overall classification accuracy was 99%, with a 99% sensitivity and specificity of 99%. The NB obtained 91% classification accuracy with 95% specificity and 94% sensitivity, SVM achieved 98% classification accuracy with a 98% specificity and sensitivity of 99%. The DT achieved 98% classification accuracy with a 98% specificity and a sensitivity of 100%. The RF achieved 100% classification accuracy with 99% specificity and 100% sensitivity.

Table 9.8. Comparison of Sensitivity and Specificity (With All (24+1) Features)

Training - Testing Partition	**Sensitivity/Specificity with all features**											
	LR		**NB**		**SVM**		**DT**		**RF**		**EDT**	
	Specificity	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**
50-50	0.97	0.99	0.94	1	0.98	0.99	0.98	1	0.97	1	1	1
66-34	0.93	0.88	0.94	1	0.95	0.98	0.99	1	0.99	1	1	1
80-20	0.92	0.93	0.98	1	0.98	0.96	1	1	1	1	1	1
10-fold Cross Validation	0.87	0.94	1	0.94	0	1	0.97	0.98	0.99	1	1	1

EDT ensures minimal errors with the dataset taken from the UCI machine learning dataset. High specificity ensures fewer patients would have to be tested for CKD without the disorder, minimizing type-2 error. Simultaneously, higher sensitivity value could save money and shorten waiting time for treatment, crucial to saving lives, minimizing type-1 error.

Table 9.9. Comparison of Sensitivity and Specificity (With Selected (13+1) Features)

Training - Testing Partition	Sensitivity/Specificity with all features											
	LR		NB		SVM		DT		RF		EDT	
	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
50-50	0.90	0.99	0.87	0.94	0.98	0.99	0.97	1	0.97	0.97	1	1
66-34	0.99	0.97	0.87	0.94	0.98	0.99	0.97	1	0.99	0.99	1	1
80-20	0.99	0.97	0.87	0.94	0.98	0.99	0.97	1	0.99	0.99	1	0.96
10-fold Cross Validation	0.95	0.97	0.95	0.88	0.79	0.75	0.98	0.98	0.99	1	0.99	1

Table 9.10. and Table 9.11. show the ROC charts using different ML techniques for these studies. Six ROC charts were drawn in various sections in blue / red color for 10-fold cross-validations. Experimental results show that our proposed classification technique's accuracy has surpassed all of the other states of the art classifiers mentioned in the literature review for ten-fold cross-validation. With our proposed classifier, the generated value of the AUC is 1.

Table 9.10. Comparison of ROC Curve and AUC Values with All (24+1) Features

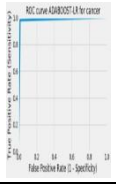
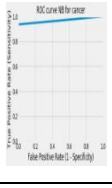
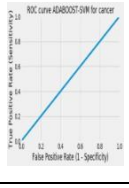
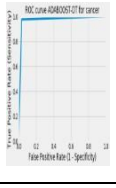
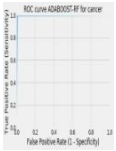
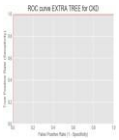
Training - Testing Partition	LR	NB	SVM	DT	RF	EDT
10-fold cross validation						
AUC	99.627	99.707	73.187	97.133	99.973	100

Table 9.11. Comparison of ROC Curve and AUC Values with All (13+1) Features

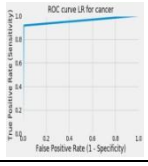
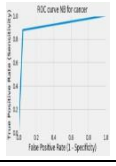
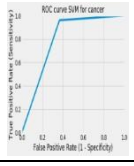
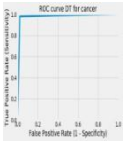
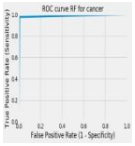
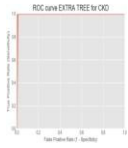
Training - Testing Partition	LR	NB	SVM	DT	RF	EDT
10-fold cross validation						
AUC	99.28	98.173	90.987	98.0	99.987	100

Table 9.12. and Table 9.13. show the K_v analysis results of the six ML techniques used in this study and our proposed model. This analysis proved that the proposed model worked significantly better than other classifiers (value = 1).

Table 9.12. Comparison of Kappa (With All (24+1) Features)

Training - Testing Partition	Kappa Statistics with All Features					
	LR	NB	SVM	DT	RF	EDT
50-50	0.95	0.92	0.97	0.98	0.96	1
66-34	0.95	0.92	0.92	0.98	0.97	1
80-20	0.97	0.97	0.95	1	0.97	1
10-fold cross validation	0.97	0.92	0.0	0.94	0.99	1

Table 9.13. Comparison of Kappa (With Selected (13+1) Features)

Training - Testing Partition	Kappa Statistics with Selected Features					
	LR	NB	SVM	DT	RF	EDT
50-50	0.85	0.78	0.96	0.96	0.94	1
66-34	0.97	0.78	0.96	0.96	0.98	1
80-20	0.97	0.78	0.96	0.96	0.98	1
10-fold cross validation	0.93	0.81	0.52	0.96	0.99	1

The study also applied a Wilcoxon signed-rank test (WSRT), a non-parametric statistical test for performing a comparative analysis of performances for various ML techniques used in this research paper. By comparing the median of a single column of numeric values against a hypothetical median, this hypothesis test evaluates the statistical differences between two populations (Wilcoxon 1945). In the present research paper, 6 class problems have solved using WSRT (Schreiner et al. 2019; Chung et al. 2017; Raghavendra et al. 2018). To handle this multi-class problem, we utilized a one-against-one approach, which breaks x number of classes into binary classes containing a set of all possible pairs of x classes. In this study, six-class classification has broken down into 15 binary sub-problems. Using the WSRT test, the significant features have evaluated separately for each binary issue. As explained in the following subsection, each binary issue's selected features were used as input to LR, RF, EDT, NB, SVM, and DT classifier.

Table 9.14. Wilcoxon Rank-Sum Test

	LR	**RF**	**EDT**	**NB**	**SVM**	**DT**
LR	0.008	0.01	0.405	0.02	0.154
RF		0.764	0.007	0.007	0.04
EDT			0.007	0.02	0.023
NB				0.05	0.019
SVM					0.765

Table 9.14. lists the p values of the WSRT for the pairs of accuracies originating from the analysis of different machine-learning algorithms performed in this research work by splitting the CKD dataset into various training and testing partitions. According to the results, there is no significant difference between LR-NB, LR-DT, RF-EDT, NB-SVM, NB-DT, and SVM-DT since the p-value is >0.05. The WSRT indicates that those algorithm pairs are mutually convergent. We cannot reject the null hypothesis for the above cases. However, our proposed model produces different results with all the other classifiers, and in all such cases, the p-value is less than 0.05 (LR-EDT, EDT-NB, EDT-SVM, EDT-DT), which means the medians of these distributions differ. Thus, the null hypothesis H0 for all these pairs can be

rejected, which indicates that our proposed model considerably outperforms the compared classification models.

Thus, the proposed method, EDT exhibits salient features, namely: **Consistency/approximation:** It is explicitly carried out by single-tree dependent techniques; **Low computational complexity:** The suggested approach's complexity is proportional to nlogn (where n is the number of sub-datasets), superior to that of single trees if the number of input variables is very high; **Robust:** The Proposed method remains robust to irrelevant and redundant variables; **Optimize the bias-variance trade-off:** The optimization of bias-variance tradeoff refers to balancing underfitting and overfitting (Wahba et al. 1994; Wahba et al. 1998; Briscoe & Feldman 2011). This method optimizes bias-variance trade-off. Its ensemble trees with data sub-sets are drawn from the dataset without replacement. As a result, the accuracy, sensitivity, and specificity are found to be of very high order; **Robustness to outliers:** On induced models, outliers tend to have a very high local effect, which makes the approach potentially much more stable than the techniques of parametric linear or non-linear least square regression. The results from this method are not affected by outliers; **Interpretability:** Tree-based models emphasize the significance of input variables in a simple way and are capable of justifying their logic.

The proposed algorithm – combining RFE and EDT enables 100 percent classification of CKD patients.

Chronic kidney disease (CKD) has features distinguishing it from non-CKD, i.e., high serum creatinine, albumin, serum urea, and pus cells.

The data set used in this study encompasses a wide range of cases – for example, patients' age ranges from two to 90 years. It has all relevant parameters to assess whether the disease is chronic or not - age, serum creatinine, albumin, urea, pus cells, specific gravity, coronary artery disease, hypertension, etc. Medical research showed the impact of different (one or more) significant factors contributing to CKD. The notable findings are that reduction of creatinine extraction indicates the disease; the presence of pus cells and albumin beyond threshold levels act as CKD warnings. Prediction of CKD is also dependent on the age of the

patient, hypertension, and specific gravity. The presence of albumin in urine has a close association with kidney failures and acts as a critical CKD marker in adults. CKD has a close association with hypertension as it causes the weakening of blood vessels and arteries, resulting in low filtering of waste in the body by the kidney. Thus, CKD is also identified with coronary artery disease (CAD). Therefore, elevated blood pressure, CAD, and hypertension are robust diagnostic features for CKD. Medical test results often show varying levels of different parameters (24), making interpretation of results difficult; this results in Type-1 or Type-2 errors. Thus, there is a need to identify significant features. Machine learning algorithms identify significant ones without affecting the sensitivity, specificity, and precision results.

The proposed algorithm performed far superior and consistent in terms of the statistical tests compared to past research outcomes. Table 9.15. provides the comparison indicating that errors are significantly less with tree-based classifiers and the classification by the proposed model was cent percent precise with AUC-ROC equal to 100.

Table 9.15. Comparison of Performance with Existing Literature

Reference	Risk Factors	ML Technique	Acc	Sen	Spe	Pre	F1	AUC	Kappa
(Alaiad et al. 2020)	12	NB	94.50	99.56	87.65	91.60			
		DT (J48)	96.75	97.21	95.97	97.60			
		SVM	97.75	100	94.34	96.40			
		KNN	98.50	99.59	96.75	98.00			
		JRip	96.00	96.42	95.27	97.20			
(Rubini & Perumal 2020)	11	FFOA, MKSVM	98.51	97.6	100				
(Almasoud & Ward 2019)	15	LR	98.75	98.4	99.33	99.5	98.9	99.7	
		SVM	97.5	96.4	99.33	99.5	97.9	99.9	
		RF	98.5	99.6	96.6	98.0	98.7	99.5	
		GDB	99.0	98.8	99.33	99.5	99.1	99.9	
(Almansour et al. 2019)	12	ANN	99.75	99.6		100	99.7		
		SVM	97.75	96.4		100	98.2		
	-	NB	99.10						

(Besra & Majhi 2019)		SMO	99.55					
		IB1	99.90					
		Multiclassifier	98.30					
		VFI	99.48					
		RF	99.43					
(Devika, Avilala, & Subramaniyaswam 2019)	24	NB	99.64	99.6		100		
		KNN	87.78	87.7		87.9		
		RF	99.84	99		99.85		
(Elhoseny, Shankar, & Uthayakumar 2019)	14	D-ACO	95.00	96.00	93.33		96.00	89.33
		ACO	87.50	88.88	84.61		90.56	72.06
		PSO	85.00	88.00	80.00		88.00	68.00
		OlexGA	75.00	80.00	66.66		80.00	46.66
(Hasan & Hasan 2019)	13	Adaptive Boosting	99	98	100	100	99	
		Bootstrap Aggregating	96	100	89	94	95	
		Extra Trees	98	97	100	100	98	
		GDB	97	98	95	97	97	
		Random Forest	95	97	97	95	95	
(Saringat et al. 2019)	24	ZeroR	62.50	100		62.50		
		Rule Induction	92.50	46.11		46.27		
		SVM	90.25	46.37		44.79		
		NB	98.50	66.00		65.40		
		DT	95.50	63.78		63.41		
		Decision Stump	92.00	89.60		97.40		
		KNN	94.75	63.91		62.73		
		Classification via Regression	98.25	97.60		99.60		
(Aljaaf et al. 2018)	7	RPART	95.6	93.39	100	100	96.5	98.2
		SVM	95.0	98.92	89.55	92.92	95.8	97.3
		LR	98.1	98.97	96.77	97.97	98.4	99.4
		MLP	98.1	98.97	96.77	97.97	98.4	99.5
(Hore et al. 2018)	23	NN	98.33	100		95.74		
		RF	92.54	96		85.71		
		MLP-FFN	99.5	100		99.2		

		NN-GA	100	100		100	
(Kemal 2018)	17	kNN	95.75				91.4
		SVM	98.25				96.3
		RBF	98.75				97.35
		Random Subspace	99.75				99.47
(Tikariha & Richhariya 2018)	24	KNN	98.5	98.5		98.5	
		SVM	97.75	97.8		97.9	
		NB	94.5	94.5		95.1	
		C4.5	96.75	96.8		96.7	
(Zeynu & Patil 2018)	8	KNN	99	99		99	
	7	J48	97.25	97.3		97.3	
	8	ANN	99.5	99.5		99.5	
	9	NB	99	99		99	
	8	SVM	98	98		98	
(Alasker et al. 2017)	8	ANN	100	100	100		
		NB	100	100	100		
		Decision Table	97.62	91.2	100		
		J48	98.41	94.1	100		
		OneR	99.21	97.1	100		
		KNN	97.62	91.2	100		
(Basar & Akan 2017)	10	Random Subspaces + REPTree	99.75				99.47
		Random Subspaces + BFTree	100				100
		Bagging + J48Tree	99.75				99.47
		Adaboost + SVM	98.5				96.81
(Chatterjee et al. 2017)	23	MLP-FFN	96.33	100		95.74	
		NN-GA	97.5	100		99.2	
		NN-CS	99.2	100		99.4	
(Wibawa, Maysanjaya, & Putra 2017)	17	CFS + AdaBoost + NB	98.0	98.0		98.1	

		CFS + AdaBoost + KNN	98.1	98.0		98.0			
		CFS + AdaBoost + SVM	97.5	97.5		97.5			
(Charleonnan et al. 2016)	24	SVM	98.3	99	98				
		LR	96.55	94	98				
		DT	94.8	93	96				
		KNN	98.1	96	99				
(Tazin, Sabab, & Chowdhury 2016)	15	NB	96						91.6
		SVM	98.5						96.8
		DT	99						97.9
		KNN	97.5						94.7
(Chetty, Vaisla, & Sudarsan 2015)	6	NB	99						
	12	SMO	98.25						
	7	IBK	100						
This Study	**13**	**EDT**	**100**	**100**	**100**	**-**	**-**	**100**	**100**

This study identified 13 significant factors – albumin, sugar, state-of red blood cells (nominal), pus cells, serum-creatinine, potassium level, hemoglobin, Count of red blood cells, hypertension, state-of diabetes mellitus, appetite, state-of pedal edema, and state-of anaemia.

The association rules that can be derived are – patients with

1. Abnormal values of the 13 factors, i.e., either very high or low, showed CKD's presence. For example, low blood count, i.e., below 4 MCL, potassium level of below 3, serum creatinine greater than 1.3 milligrams per deciliter with pus cell, diabetes, aenemia, pedal edema, and hypertension is a definite case of CKD.

2. Albumin level >= 1 have always shown signs of CKD under different combinations of other factors – Appendix 1

The most critical factors observed are - low levels of haemoglobin, RBC, potassium, and high creatinine and sugar levels.

3. Albumin level = 0 can have CKD (56 cases reported in the data set)

if serum-creatinine >= 0.5 and RBC = Abnormal or Anemic or hemoglobin <= 0.5

4. Pedal Edema = 1, a nephrotic syndrome due to CKD in cases where:

Serum creatinine is between 32 and 0.5, potassium levels lie between 39 and 2.5, and RBC-count is between 5.4 and 2.1, with or without pus cell, sugar, and hypertension. Thus, a patient with nephrotic syndrome and low - potassium levels, RBC-count and haemoglobin, and or high creatinine, hypertension, and sugar are likely to suffer from CKD. This data set showed 78 cases with such instances.

The outcome of EDT (under 50 - 50 split) indicates albumin, red blood cells, Diabetes Mellitus, Haemoglobin, Hypertension, and Red Blood Cell Count as significant compared to other six features (Pus Cell, Pedal Edema, Serum Creatinine, Potassium, Appetite, Sugar and Anemia). This is reflected in Figure 9.5.. The results delineate the causes and effects of CKD. The features found significant are causes of CKD (as evident from the literature) while the other features are effects of CKD, namely low extraction of creatinine, presence of pus-cells and potassium, low appetite, and anaemia. Similar results were also obtained for other split ratios including 10-fold cross-validation.

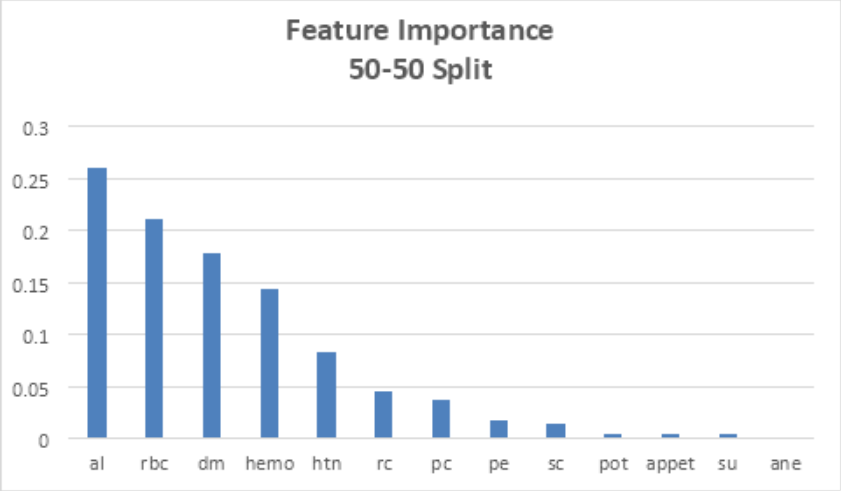


Figure 9.5. Feature Importance (50-50 Split) Obtained from EDT Classifier

The algorithm showed a steady behaviour when all features were taken. However, the sensitivity and specificity were less than 1 in two cases when the significant features were selected and under different test-train split ratios. These corroborate the results shown in table

The sensitivity was found to be 0.96 under 80 – 20 split, and specificity was 0.99 under 10-fold cross validation.

However, this behaviour can be accepted as the composite ROC-AUC score was one. This metric has proven to be meaningful for comparing models and providing more clarity than other metrics, namely accuracy, sensitivity, and specificity (Vandewiele et al. 2020).

An i5 laptop with 4 GB DDR4 RAM and 64-bit Windows 10 operating system was used for computational work. The computation for different cases took time less than a second, except in the case of 10-fold cross-validation, where the time was around 38 seconds. The time was calculated using the "timeit" function from the Python library.

Conclusion

Chronic diseases such as CKD decrease patients' working life due to frequent visits to doctors, hospitalization, and anxiety. So far, several machine learning techniques have been applied to study and diagnose diseases. However, previous studies missed out on segregating the disease as chronic or non-chronic, the significant features that affect chronic diseases, and the stress on sensitivity and specificity (and not only accuracy). This research paper proposes an algorithm that classifies CKD and non-CKD patients, identifies the significant features using the proven techniques, and determine the best combination of the features that provide the higher sensitivity and specificity in results. The results from the proposed method were more accurate than other standard techniques. The ten-fold cross-validation technique with moderate iterations yields the highest accuracy levels than any other study so far. Thus, this paper stands out from all studies done so far: the association rules and feature importance obtained in this exercise aids in effective diagnosis and treatment of CKD.

The proposed method aims at optimizing the bias/variance tradeoff to balance underfitting and overfitting (Wahba et al. 1994; Wahba et al. 1998; Briscoe & Feldman 2011). However,

this approach increases variance. Unlike other regression methods, the growth of trees is quite adaptive to the training samples, which means that the induced models are very different for different samples of the same size taken from the same distribution. This variation usually translates to a low accuracy, which is undoubtedly the system's most significant drawback. Tree pruning is typically unable to significantly reduce this variance while optimizing the bias/variance tradeoff.

Chapter 10

Conclusions and Future Work

In this thesis, we successfully make use of machine learning techniques to solve some problems arising from biological and clinical data. We have articulated explicitly the 2-step framework of feature selection, and feature integration with learning algorithms and demonstrated the effectiveness of ensemble learning when dealing with classification and patient survival prediction from patient datasets.

From a large number of experiments conducted on some high-dimensional medical data sets, we observe the improvements in performances of all the classification algorithms under the proposed feature selection scenarios and stacking of classifiers.

In general, features selections methods and stacking of classifiers can improve the performance of ML algorithms. However, no single features selections method that best satisfies all datasets and learning algorithms. Therefore, machine learning researchers should understand the nature of datasets and learning algorithm characteristics to obtain better outcomes as possible.

In the aspect of classification algorithms, no single algorithm is superior to all others, though some of the algorithms achieve fairly good results in most tests. Compared with others, decision tree methods can provide simple, comprehensive rules and are not very sensitive to feature selections.

The first application of ML techniques in this thesis finds the right age for persons suffering from CVDs. Literature review shows that several DMTs of varied accuracy level exists. The above study shows that four variables (namely age, cp, ca, and thal) typically affect CVD. The insurance companies fix the premiums based on these four variables. The slabs of the premium will vary according to the probability of CVD based on different values of these four variables. The techniques namely Decision tree classifiers, K-Means, and Logistic Regression have been proposed to derive the complete set of association rules leading to the

identification of CVDs. Using any one of those methods is likely to contribute to the Type 1 error and none have a 100% accuracy level.

Decision tree analysis shows that factors namely thal, ca and age are significant to predict the probability of CVD. K-Means classifiers further show that the age group for CVD can be lower if a partial variable set is considered. The logistic regression model indicates the probability of CVD as the sum of the product of these variables namely cp, resting, ca, and thal with their coefficients 0.6219, 0.4466, 0.9547, and .4649 respectively. That is, logistic regression establishes the weightage of the significant factors. However, a cluster of variables with specific values defines the possibility of CVD. Therefore, the significant factors need to be identified, their weightage determined and clustering with age is determined for customizations of insurance premiums. Therefore, the association rule can be established accordingly. Such findings can be the basis for customizing insurance premiums for both the categories of life and health instead of the traditional system of calculating premiums on age-wise slabs.

The use of the K-means Cluster approach suggests that the consideration of partial variable set reveal persons with age less than 40 can also suffer from CVD. This is true when the interactive effect of three variables namely sex, ca and trestbps are considered in place of all parameters. The decision tree further lays down the association rule strengthening the findings of K-Means analysis. The results of the decision tree show that the existence of all three modifiable risk factors significantly increases the probability of CVD.

However, since the observations from K-Means analysis are more precise when age is focused, its results are included in the association rule. The Logistic Regression analysis shows that the variables such as cp, restecg, ca, and thal with definite weightage can lead to CVD. Thus, it can be concluded that instead of using any one technique, the combination of the most reliable techniques as listed in this study can be used to define the association rule.

In the second application of ML technique, various ML techniques were used with varying degrees of accuracy. The reasons, proposed by authors, for this variedness include the difference in the dataset, imbalance in the dataset, and capability of the classifiers. Each

classifier has its own merits and demerits, and none of the popular ones could establish complete superiority over all others. The authors in this application attempted to make use of the existing techniques and determine whether the application of techniques on dataset under two conditions i.e., – i. containing the significant variables alone and dataset; ii. containing the non-contradicting variables could enhance the accuracy of the prediction. Logistic Regression analysis without contradicting variables is showed greater accuracy levels in non-recurrence of breast cancer. Thus, the paper contributed to identifying the prediction with higher accuracy by an ensemble of outcomes of individual approaches. The degree of malignancy was found to the most significant variable, as the cause of recurrence of breast cancer, as it appeared significant in all forms of analysis.

Thus, the approach of integration of data mining approaches and use of relatively significant variables, and the elimination of contradicting variables lead to better prediction.

In the third application of ML techniques, the analysis of the cervical cancer disease data set highlights the cause of the disease. The common factors across the different approaches include the age of the patient, assuming all of them had multiple sex partners and used contraceptives, and suffered from sexually transmitted diseases. The results from K-means clustering show that age equal to 24 years with first sexual interaction at the lowest age of 16 caused cervical cancer.

The treatment dataset indicates the treatment attributes once the warts are detected. Here, time is a critical factor. The results from K-means clustering show that most likely time equal to around seven months with an age of 32 years showed signs of cancer. The minimum time of occurrence of cervical cancer stands out as three months post detection of warts. There are four cases where cancer got detected after six months. Thus, the time range of 3 to 7 months is most crucial.

The accuracy of the prediction of the disease appeared to be highest with the approach, namely Decision Tree. The probability of type 1 and 2 errors was 12.7% and 3.2% respectively. The association rules associated with this approach are:

Schiller's test result factor is the best factor for the detection of cervical cancer.
For the Schiller's test result = 1, 64.9 % of women have cervical cancer.
For the Schiller's test result = 0, next best predictor is age
If age > 19 and <= 21 then, 6 % of women would suffer from cervical cancer
For the age <= 19, next best predictor is STDs : pelvic inflammatory disease
For STDs : pelvic = 0.0, 0.0 % women suffer from cervical cancer
For STDs : pelvic = missing, 3.1 of women suffer from cervical cancer
For the age > 21, next best predictor is First sexual intercourse
For First sexual intercourse <= 14, 2.6% of women suffer from cervical cancer
For First sexual intercourse > 14, 0.0% of women suffer from cervical cancer

The accuracy of the prediction of the treatment of the disease appeared to be highest with approaches, namely Logistics-Regression and Decision Tree. The Decision-Tree approach yielded a 7% type 1 error. The association rules associated with these two approaches are:

For the time <= 5.250, 84.6 % of women get treated.
If Time > 5.250 and <= 8.000 then, 100% of women get treated
If Time > 8.000 and <= 10.500 then, 83.3% of women get treated
For the time > 10.500, 29.4 % of women get treated.

Decision Tree analysis showed consistency in performance, while Random-forest performed better in a data set with a distinct dichotomy. In the case of disease data, the ratio of biopsy = 1 and biopsy = 0 was 94:6. Hence, the results from the RF-based Decision Tree were not significant. Whereas in the treatment dataset, the ratio was 80:20, the results reflected higher prediction accuracy. The accuracy level of the LR on RF determined variables showed a superior result (96.2%) among all methods, for predicting the treatment.

Hence, the alternate hypothesis – "integration of data mining approaches leads to better prediction" stands accepted for a dataset with dichotomy at the ratio of 80:20 or, better. The significant variables when the disease is detected include – Age, schiller – STD parameters. The critical factor for successful treatment is Time. The critical values of these factors are – Age = 19 years and above; Time = 3 to 7 months; Schiller Test = 1; STD test (of different parameters) = 1.

In the fourth application in this research work, different data mining classification techniques were used for the prediction of various chronic and recurrent diseases and their performance was compared to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. It has been noticed that an ensemble of classifiers and repetitive over-sampling may improve the performance of dealing with class-imbalanced problems.

The author has professed various methods with varying degrees of accuracy for the classification of different datasets. The reasons, proposed by authors, for this variedness include the difference in the dataset, imbalance in the dataset, and capability of the classifiers. Each classifier has its own merits and demerits, and none of the popular ones could establish complete superiority over all others.

The fifth application considers Chronic Disease. Chronic diseases such as CKD decrease patients' working life due to frequent visits to doctors, hospitalization, and anxiety. So far, several machine learning techniques have been applied to study and diagnose diseases. However, previous studies missed out on segregating the disease as chronic or non-chronic, the significant features that affect chronic diseases, and the stress on sensitivity and specificity (and not only accuracy). In this paper, the authors propose an algorithm that classifies CKD and non-CKD patients, identifies the significant features using the proven techniques and determines the best combination of the features that provide the higher sensitivity and specificity in results. The paper introduces the stacked classification technique – EDT for CKD diagnosis and found it to be accurate than other standard techniques. The ten-fold cross-validation technique with moderate iterations yields the highest accuracy levels than any other study so far. Thus, this paper stands out from all studies done so far.

The results, especially related to identifying the significant features, can be further validated by considering other datasets. This limitation can be overcome in future research work.

References

Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., & Gururajan, R. (2020). A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters, 132, 123-131.

Abed, M., & Ibrikci, T. (2019, September). Comparison between Machine Learning Algorithms in the Predicting the Onset of Diabetes. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-5).

Afzal, Z., Schuemie, M. J., van Blijderveen, J. C., Sen, E. F., Sturkenboom, M. C., & Kors, J. A. (2013). Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC medical informatics and decision making, 13(1), 1-11.

Ahmad, M., Tundjungsari, V., Widianti, D., Amalia, P., & Rachmawati, U. A. (2017, November). Diagnostic decision support system of chronic kidney disease using support vector machine. In 2017 Second International Conference on Informatics and Computing (ICIC) (pp. 1-4). IEEE.

Ahmed, B., Thesen, T., Blackmon, K. E., Kuzniekcy, R., Devinsky, O., & Brodley, C. E. (2016). Decrypting "Cryptogenic" Epilepsy: Semi-supervised Hierarchical Conditional Random Fields For Detecting Cortical Lesions In MRI-Negative Patients. Journal of Machine Learning Research, 17(112), 1-30.

Ahmedin Jemal, D. A., Tiwari, R. C., Murray, T., Ghafoor, A., Samuels, A., Ward, E., ... & Thun, M. J. (2004). Cancer statistics. 2004. CA Cancer J Clin, 54(1), 8-29.

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2013). Good practice in large-scale learning for image classification. IEEE Trans. Pattern Anal. Mach. Intell., 36(3), 507-520.

Akgül, M., Sönmez, Ö. E., & Özcan, T. (2019, July). Diagnosis of Heart Disease Using an Intelligent Method: A Hybrid ANN–GA Approach. In International Conference on Intelligent and Fuzzy Systems (pp. 1250-1257). Springer, Cham.

Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). Classification and association rule mining technique for predicting chronic kidney disease. Journal of Information & Knowledge Management, 19(01), 2040015.

Alaoui, S. S., Aksasse, B., & Farhaoui, Y. (2018, July). Statistical and Predictive Analytics of Chronic Kidney Disease. In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 27-38). Springer, Cham.

Alasker, H., Alharkan, S., Alharkan, W., Zaki, A., & Riza, L. S. (2017, October). Detection of kidney disease using various intelligent classifiers. In 2017 3rd International Conference on Science in Information Technology (ICSITech) (pp. 681-684). IEEE.

Alassaf, R. A., Alsulaim, K. A., Alroomi, N. Y., Alsharif, N. S., Aljubeir, M. F., Olatunji, S. O., Alahmadi, A. Y., Imran, M., Alzahrani, R. A., & Alturayeif, N. S. (2018, November). Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques. In 2018 International Conference on Innovations in Information Technology (IIT) (pp. 99-104). IEEE.

Albalate, A., & Minker, W. (2013). Semi-supervised and unsupervised machine learning: novel strategies. John Wiley & Sons.

Alderman, A. K., Wilkins, E. G., Lowery, J. C., Kim, M., & Davis, J. A. (2000). Determinants of patient satisfaction in postmastectomy breast reconstruction. Plastic and reconstructive surgery, 106(4), 769-776.

Alebiosu, C. O., & Ayodele, O. E. (2005). The global burden of chronic kidney disease and the way forward. Ethn Dis, 15(3), 418.

Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on X2 Statistical Model and Optimally Configured Deep Neural Network. IEEE Access, 7, 34938-34945.

Aličković, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4), 753-763.

Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2018, July). Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-9). IEEE.

Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P., & Aljaaf, A. J. (2020). Performance-Based Prediction of Chronic Kidney Disease Using Machine Learning for High-Risk Cardiovascular Disease Patients. In Nature-Inspired Computation in Data Mining and Machine Learning (pp. 187-206). Springer, Cham.

Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Comput. Biol. Med., 109, 101-111.

Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Comput Sci Appl, 10(8).

Al-Shamsi, S., Regmi, D., & Govender, R. D. (2018). Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: a population-based study. PloS one, 13(6), e0199920.

Amarbayasgalan, T., Van Huy, P., & Ryu, K. H. (2020). Comparison of the Framingham Risk Score and Deep Neural Network-Based Coronary Heart Disease Risk Prediction. In Advances in Intelligent Information Hiding and Multimedia Signal Processing (pp. 273-280). Springer, Singapore.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A. and Hussain, A., 2016. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4, pp.7940-7957.

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telemat. Inform., 36, 82-93.

Ananian, P., Houvenaeghel, G., Protiere, C., Rouanet, P., Arnaud, S., Moatti, J. P., ... & Julian-Reynier, C. (2004). Determinants of patients' choice of reconstruction with mastectomy for primary breast cancer. Annals of Surgical Oncology, 11(8), 762-771.

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Meth Prog Bio, 141, 19-26.

Arai, H., Maung, C., Xu, K., & Schweitzer, H. (2016, February). Unsupervised feature selection by heuristic search with provable bounds on suboptimality. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., & He, Y. L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. Inf. Sci., 378, 484-497.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

Atlas, L., Connor, J., Park, D., El-Sharkawi, M., Marks, R., Lippman, A., ... & Muthusamy, Y. (1989, November). A performance comparison of trained multilayer perceptrons and trained classification trees. In Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics (pp. 915-920). IEEE.

Ayat, N. E., Cheriet, M., & Suen, C. Y. (2005). Automatic model selection for the optimization of SVM kernels. Pattern Recognition, 38(10), 1733-1745.

Aydın, E. A., & Kaya Keleş, M. (2017). Breast cancer detection using K-nearest neighbors data mining method obtained from the bow-tie antenna dataset. Int J RF Microw C E, 27(6), e21098.

Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. IETE J Res, 1-20.

Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. Comput Methods Programs Biomed, 113(2), 465-473.

Baati, K., Hamdani, T. M., Alimi, A. M., & Abraham, A. (2016). A new possibilistic classifier for heart disease detection from heterogeneous medical data. Int J Comput Sci Inform Secur, 14(7), 443.

Bani-Hani, D., Patel, P., & Alshaikh, T. (2019). An Optimized Recursive General Regression Neural Network Oracle for the Prediction and Diagnosis of Diabetes. Global Journal of Computer Science and Technology, 19(2-D).

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE T Inf Technol B, 14(4), 1114-1120.

Basar, M. D., & Akan, A. (2017). Detection of chronic kidney disease by using ensemble classifiers. In 2017 10th International Conference on Electrical and Electronics Engineering (ELECO) (pp. 544-547). IEEE.

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving Heart Disease Prediction Using Feature Selection Approaches. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 619-623). IEEE.

Beard, J. R., Officer, A., De Carvalho, I. A., Sadana, R., Pot, A. M., Michel, J. P., ... & Thiyagarajan, J. A. (2016). The World report on ageing and health: a policy framework for healthy ageing. The lancet, 387(10033), 2145-2154.

Besra, B., & Majhi, B. (2019). An Analysis on Chronic Kidney Disease Prediction System: Cleaning, Preprocessing, and Effective Classification of Data. In Recent Findings in Intelligent Computing Techniques (pp. 473-480). Springer, Singapore.

Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 42(10), 4611-4620.

Bhatia, K., & Syal, R. (2017, October). Predictive analysis using hybrid clustering in diabetes diagnosis. In 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE) (pp. 447-452). IEEE.

Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), 1112.

Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Medicine, 9(9), 3234-3243.

Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Information Sciences, 282, 111-135.

Borisagar, N., Barad, D., & Raval, P. (2017). Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm. In Proceedings of International Conference on Communication and Networks (pp. 295-303). Springer, Singapore.

Breiman, L. (1996). Bagging predictors. Mach Learn, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432, 151-166.

BRFSS. https://www.cdc.gov/brfss/annual_data/annual_ 2014.html. Accessed 28 June, 2020.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. Cognition, 118(1), 2-16.

Burse, K., Kirar, V. P. S., Burse, A., & Burse, R. (2019). Various preprocessing methods for neural network based heart disease prediction. In Smart innovations in communication and computational sciences (pp. 55-65). Springer, Singapore.

Cai, Q., K Mukku, V., & Ahmad, M. (2013). Coronary artery disease in patients with chronic kidney disease: a clinical update. Current cardiology reviews, 9(4), 331-339.

Cai, Y., Tan, X., & Tan, X. (2017). Selective weakly supervised human detection under arbitrary poses. Pattern Recognit, 65, 223-237.

Canlas, R. D. (2009). Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carnegie Mellon University, Australia.

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., ... & Witmer, J. (2013). STAT2: building models for a world of data.

Cao, D. S., Xu, Q. S., Liang, Y. Z., Chen, X., & Li, H. D. (2010). Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. Chemometrics and Intelligent Laboratory Systems, 103(2), 129-136.

Castro, M. D. F., Mateus, R., Serôdio, F., & Bragança, L. (2015). Development of benchmarks for operating costs and resources consumption to be used in healthcare building sustainability assessment methods. Sustainability, 7(10), 13222-13248.

Cenamor, I., de la Rosa, T., Núñez, S., & Borrajo, D. (2017). Planning for tourism routes using social networks. Expert Systems with Applications, 69, 1-9.

Chalak, L. F., Pavageau, L., Huet, B., & Hynan, L. (2020). Statistical rigor and kappa considerations: which, when and clinical context matters. Pediatric research, 88(1), 5-5.

Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. (2016, October). Predictive analytics for chronic kidney disease using machine learning techniques. In 2016 Management and Innovation Technology International Conference (MITicon) (pp. MIT-80). IEEE.

Chatterjee, S., Banerjee, S., Basu, P., Debnath, M., & Sen, S. (2017, April). Cuckoo search coupled artificial neural network in detection of chronic kidney disease. In 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech) (pp. 1-4). IEEE.

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 12(2), 119-126.

Chen, K., Zhou, F. Y., & Yuan, X. F. (2019). Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. Expert Systems with Applications, 128, 140-156.

Chen, W., Chen, S., Zhang, H., & Wu, T. (2017, November). A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 386-390). IEEE.

Chen, Z., Zhang, Z., Zhu, R., Xiang, Y., & Harrington, P. B. (2016). Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers. Chemometr Intell Lab Syst, 153, 140-145.

Chetty, N., Vaisla, K. S., & Sudarsan, S. D. (2015, December). Role of attributes selection in classification of Chronic Kidney Disease patients. In 2015 International Conference on Computing, Communication and Security (ICCCS) (pp. 1-6). IEEE.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc, 24(2), 361-370.

Choubey, D. K., Kumar, P., Tripathi, S., & Kumar, S. (2020). Performance evaluation of classification methods with PCA and PSO for diabetes. Netw Model Anal Health Inform Bioinform, 9(1), 5.

Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2017, February). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016) (pp. 451-455).

Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In Recent Developments in Machine Learning and Data Analytics (pp. 67-78). Springer, Singapore.

Chronic Kidney Disease Prognosis Consortium. (2010). Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. The Lancet, 375(9731), 2073-2081.

Chui, K. T., Alhalabi, W., Pang, S. S. H., Pablos, P. O. D., Liu, R. W., & Zhao, M. (2017). Disease diagnosis in smart healthcare: Innovation, technologies and applications. Sustainability, 9(12), 2309.

Chung, C. J., Kuo, Y. C., Hsieh, Y. Y., Li, T. C., Lin, C. C., Liang, W. M., ... & Lin, H. C. (2017). Subject-enabled analytics model on measurement statistics in health risk expert system for public health informatics. International journal of medical informatics, 107, 18-29.

CKD-A. Rubini, L. J., Eswaran, P., & Soundarapandian, P. (2015). UCI Chronic Kidney Disease. School of Information and computer Sciences, University of California, Irvine, CA. https://archive.ics.uci.edu/ml/ datasets/Chronic_Kidney_Disease. Accessed 26 June, 2020.

Coresh, J., Wei, G. L., McQuillan, G., Brancati, F. L., Levey, A. S., Jones, C., & Klag, M. J. (2001). Prevalence of high blood pressure and elevated serum creatinine level in the United States: findings from the third National Health and Nutrition Examination Survey (1988-1994). Archives of internal medicine, 161(9), 1207-1216.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cramer, D. (2003). Advanced quantitative data analysis. McGraw-Hill Education (UK).

Cvetković, B., Kaluža, B., Gams, M., & Luštrek, M. (2015). Adapting activity recognition to a person with Multi-Classifier Adaptive Training. J Ambient Intell Smart Environ, 7(2), 171-185.

Dangare, C., & Apte, S. (2012). A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology (IJCET), 3(3).

Dash, N. K. (2005). Module: Selection of the research paradigm and methodology. Retrieved August, 9, 2009.

Défossez, A., & Bach, F. (2015, February). Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In Artificial Intelligence and Statistics (pp. 205-213).

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back-propagation neural network versus logistic regression in heart disease classification. In Advanced computing and communication technologies (pp. 133-144). Springer, Singapore.

Devi, R. D. H., & Devi, M. I. (2016). Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. Int. J. Adv. Eng. Technol, 12, 93-98.

Devi, R. D. H., Bai, A., & Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obes. Med., 17, 100152.

Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 679-684). IEEE.

Dey, S. K., Hossain, A., & Rahman, M. M. (2018, December). Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT) (pp. 1-5). IEEE.

Di Noia, T., Ostuni, V. C., Pesce, F., Binetti, G., Naso, D., Schena, F. P., & Di Sciascio, E. (2013). An end stage kidney disease predictor based on an artificial neural networks ensemble. Expert systems with applications, 40(11), 4438-4445.

Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing, 54, 183-199.

Dimitoglou, J. A. Adams, and Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:1206.1121.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

Du, G., & Sun, C. (2015). Location planning problem of service centers for sustainable home healthcare: Evidence from the empirical analysis of Shanghai. Sustainability, 7(12), 15812-15832.

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural. Comput. Appl., 29(10), 685-693.

Dwivedi, A. K. (2018a). Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural. Comput. Appl., 30(12), 3837-3845.

Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease. Sci. Rep., 9(1), 1-14.

Emami, N., & Pakzad, A. (2019). A New Knowledge-Based System for Diagnosis of Breast Cancer by a combination of the Affinity Propagation and Firefly Algorithms. Journal of AI and Data Mining, 7(1), 59-68.

Erkaymaz, O., Ozer, M., & Perc, M. (2017). Performance of small-world feedforward neural networks for the diagnosis of diabetes. Appl. Math. Comput., 311, 22-28.

Escanilla, N. S., Hellerstein, L., Kleiman, R., Kuang, Z., Shull, J., & Page, D. (2018, December). Recursive feature elimination by sensitivity testing. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 40-47). IEEE.

Fahmy, T., & Aubry, A. (1998). XLstat. Société Addinsoft SARL, 40.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine, 8(1), 19-20.

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

Fauci, A. S. (Ed.). (2008). Harrison's principles of internal medicine (Vol. 2, pp. 1612-1615). New York: McGraw-Hill, Medical Publishing Division.

FHS. Framingham Heart Study Dataset. https://framinghamheartstudy.org/. Accessed 20 June, 2020.

Fisher, B., Anderson, S., Redmond, C. K., Wolmark, N., Wickerham, D. L., & Cronin, W. M. (1995). Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer. New England Journal of Medicine, 333(22), 1456-1461.

Fong, A., Clark, L., Cheng, T., Franklin, E., Fernandez, N., Ratwani, R., & Parker, S. H. (2017). Identifying influential individuals on intensive care units: using cluster analysis to explore culture. J. Nurs. Manag., 25(5), 384-391.

Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, MenloPark, C.A, 1996.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Gabbey. A. E., Jacquelyn. C. Human Papillomavirus Infection Medically reviewed by Debra Rose Wilson, PhD, MSN, RN, IBCLC, AHN-BC, CHT, 2017.

Galliers, R. D. (1992). Choosing information systems research approaches.

Gansevoort, R. T., Matsushita, K., Van Der Velde, M., Astor, B. C., Woodward, M., Levey, A. S., ... & Coresh, J. (2011). Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. Kidney international, 80(1), 93-104.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., & Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput Biol Med, 59, 125-133.

Gbengaa, D. E., Hemanthb, J., Chiromac, H., & Muhammad, S. I. (2019, January). Non-Nested Generalisation (NNGE) Algorithm For Efficient and Early Detection of Diabetes. In Information Technology and Intelligent Transportation Systems: Proceedings of the 3rd International Conference on Information Technology and Intelligent Transportation Systems (ITITS 2018) Xi'an, China, September 15-16, 2018 (Vol. 314, p. 233). IOS Press.

Gerber, B., Krause, A., Dieterich, M., Kundt, G., & Reimer, T. (2009). The oncological safety of skin sparing mastectomy with conservation of the nipple-areola complex and autologous reconstruction: an extended follow-up study. Annals of surgery, 249(3), 461-468.

Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., & Feng, M. (2015, January). A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In Aaai Conference on Artificial Intelligence (Vol. 2015, pp. 446-453).

Giveki, D., & Rastegar, H. (2019). Designing a new radial basis function neural network by harmony search for diabetes diagnosis. Opt Mem Neural Network, 28(4), 321-331.

Golberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addion wesley, 1989(102), 36.

Goyal, K., Aggarwal, P., & Kumar, M. (2020). Prediction of Breast Cancer Recurrence: A Machine Learning Approach. In Computational Intelligence in Data Mining (pp. 101-113). Springer, Singapore.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences, 84(13), 4355-4358.

Gu, B., Sheng, V. S., Tay, K. Y., Romano, W., & Li, S. (2014). Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst, 26(7), 1403-1416.

Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 291-296). IEEE.

Gunasundari, S., Janakiraman, S., & Meenambal, S. (2016). Velocity bounded boolean particle swarm optimization for improved feature selection in liver and kidney disease diagnosis. Expert Systems with Applications, 56, 28-47.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

Hahne, J. M., Biessmann, F., Jiang, N., Rehbaum, H., Farina, D., Meinecke, F. C., ... & Parra, L. C. (2014). Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control. IEEE Trans. Neural Syst. Rehabilitation Eng., 22(2), 269-279.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Haque, S. A., Rahman, M., & Aziz, S. M. (2015). Sensor anomaly detection in wireless sensor networks for healthcare. Sensors, 15(4), 8764-8786.

Hasan, K. Z., & Hasan, M. Z. (2019). Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease. In Emerging Research in Computing, Information, Communication and Applications (pp. 415-426). Springer, Singapore.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hayashi, Y., & Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Inform. Med. Unlocked, 2, 92-104.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Hore, S., Chatterjee, S., Shaw, R. K., Dey, N., & Virmani, J. (2018). Detection of chronic kidney disease: A NN-GA-based approach. In Nature Inspired Computing (pp. 109-115). Springer, Singapore.

Howlader, N., Noone, A. M., Krapcho, M., Garshell, J., Miller, D., & Altekruse, S. F. (2015). Surveillance, epidemiology, and end results (SEER) program (www. seer. cancer. gov) SEER* Stat database: incidence–SEER 18 regs research data+ hurricane katrina impacted Louisiana cases, Nov 2015 sub (2000-2013) <Katrina/Rita population adjustment>–Linke. Rita population adjustment> e Linke.

Hsu, W. C., Lin, L. F., Chou, C. W., Hsiao, Y. T., & Liu, Y. H. (2017). EEG classification of imaginary lower limb stepping movements based on fuzzy support vector machine with kernel-induced membership function. Int. J. Fuzzy Syst., 19(2), 566-579.

Huang, Z., Dong, W., Ji, L., Yin, L., & Duan, H. (2015). On local anomaly detection and analysis for clinical pathways. Artif. Intell. Med., 65(3), 167-177.

Husain, A., & Khan, M. H. (2018, April). Early diabetes prediction using voting based ensemble learning. In International Conference on Advances in Computing and Data Sciences (pp. 95-103). Springer, Singapore.

Ibrahim, A. O., & Shamsuddin, S. M. (2018). Intelligent breast cancer diagnosis based on enhanced Pareto optimal and multilayer perceptron neural network. International Journal of Computer Aided Engineering and Technology, 10(5), 543-556.

Ichikawa, D., Saito, T., Ujita, W., & Oyama, H. (2016). How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. J Biomed Inform, 64, 20-24.

Indridason, O. S., Thorsteinsdóttir, I., & Pálsson, R. (2007). Advances in detection, evaluation and management of chronic kidney disease. Laeknabladid, 93(3), 201-207.

Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 226-229). IEEE.

Jacobson, J. A., Danforth, D. N., Cowan, K. H., d'Angelo, T., Steinberg, S. M., Pierce, L., ... & Okunieff, P. (1995). Ten-year results of a comparison of conservation with mastectomy in the treatment of stage I and II breast cancer. New England Journal of Medicine, 332(14), 907-911.

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017, September). An expert system for diabetes prediction using auto tuned multi-layer perceptron. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 722-728). IEEE.

James, M. T., Hemmelgarn, B. R., Wiebe, N., Pannu, N., Manns, B. J., Klarenbach, S. W., ... & Alberta Kidney Disease Network. (2010). Glomerular filtration rate, proteinuria, and the incidence and consequences of acute kidney injury: a cohort study. The Lancet, 376(9758), 2096-2103.

Jayashree, J., & Kumar, S. A. (2019). Hybrid swarm intelligent redundancy relevance (RR) with convolution trained compositional pattern neural network expert system for diagnosis of diabetes. Health Technol (Berl), 1-10.

Jerlin Rubini, L., & Perumal, E. (2020). Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. International Journal of Imaging Systems and Technology, 30(3), 660-673.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., ... & Yang, C. W. (2013). Chronic kidney disease: global dimension and perspectives. The Lancet, 382(9888), 260-272.

Jin, L., Xue, Y., Li, Q., & Feng, L. (2016, April). Integrating human mobility and social media for adolescent psychological stress detection. In International Conference on Database Systems for Advanced Applications (pp. 367-382). Springer, Cham.

Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. Procedia computer science, 72, 306-313.

Kadam, V. J., Jadhav, S. M., & Vijayakumar, K. (2019). Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. Journal of medical systems, 43(8), 263.

Kadri, F., Harrou, F., Chaabane, S., Sun, Y., & Tahon, C. (2016). Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems. Neurocomputing, 173, 2102-2114.

Kaggle FHS. Framingham Heart study dataset. https://kaggle.com/amanajmera1/framingham-heart-study-dataset. Accessed 21 June, 2020.

Kalyankar, M. A., & Chopde, N. R. (2013). Cancer Detection: Survey. Int. Journal of Advanced Research in Computer Science and Software Engineering, 3(11), 1536-1539.

Kamadi, V. V., Allam, A. R., & Thummala, S. M. (2016). A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. Appl. Soft Comput., 49, 137-145.

Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput. Sci., 47, 45-51.

Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S. J., & Yu, K. S. (2015). An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Syst. Appl., 42(9), 4265-4273.

Kanimozhi, U., Ganapathy, S., Manjula, D., & Kannan, A. (2019). An intelligent risk prediction system for breast cancer using fuzzy temporal rules. National Academy Science Letters, 42(3), 227-232.

Kanimozhi, V. A., & Karthikeyan, T. (2016). A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease. Int. J. Adv. Res. Comput. Commun. Eng, 5(4), 552-557.

Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In Soft Computing and Medical Bioinformatics (pp. 63-72). Springer, Singapore.

Karegowda, A. G., Manjunath, A. S., &Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271-277.

Kaur, H., & Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform..

Kemal, A. D. E. M. (2018). Diagnosis of Chronic Kidney Disease using Random Subspace Method with Particle Swarm Optimization. International Journal of Engineering Research and Development, 10(3), 1-5.

Khalkhaali, H. R., Hajizadeh, I., Kazemnezhad, A., & Moghadam, A. G. (2010). Prediction of kidney failure in patients with chronic renal transplant dysfunction. Iran. J. Epidemiology, 6(2), 25-31.

Khan, F. A., Haldar, N. A. H., Ali, A., Iftikhar, M., Zia, T. A., & Zomaya, A. Y. (2017). A continuous change detection mechanism to identify anomalies in ECG signals for WBAN-based healthcare environments. IEEE Access, 5, 13531-13544.

Khourdifi, Y., & Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering & Systems, 12(1).

Kiezun, A., Lee, I. T. A., & Shomron, N. (2009). Evaluation of optimization techniques for variable selection in logistic regression applied to diagnosis of myocardial infarction. Bioinformation, 3(7), 311.

Kittler, J. (1986). Feature selection and extraction. Handbook of pattern recognition and image processing.

KNHANES. Kweon, S., Kim, Y., Jang, M. J., Kim, Y., Kim, K., Choi, S., ... & Oh, K. (2014). Data resource profile: The Korea national health and nutrition examination survey (KNHANES). Int. J. Epidemiol., 43(1), 69-77.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1006-1010). IEEE.

Kononenko, I. (1994, April). Estimating attributes: Analysis and extensions of RELIEF. In European conference on machine learning (pp. 171-182). Springer, Berlin, Heidelberg.

Koutcher, L., Ballangrud, Å., Cordeiro, P. G., McCormick, B., Hunt, M., Van Zee, K. J., ... & Beal, K. (2010). Postmastectomy intensity modulated radiation therapy following immediate expander-implant reconstruction. Radiotherapy and Oncology, 94(3), 319-323.

Koutsky, L. (1997). Epidemiology of genital human papillomavirus infection. The American journal of medicine, 102(5), 3-8.

Krijestorac, M., Halilovic, A., & Kevric, J. (2019, June). The Impact of Predictor Variables for Detection of Diabetes Mellitus Type-2 for Pima Indians. In International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (pp. 388-405). Springer, Cham.

Kriplani, H., Patel, B., & Roy, S. (2019). Prediction of chronic kidney diseases using deep artificial neural network technique. In Computer Aided Intervention and Diagnostics in Clinical and Medical Images (pp. 179-187). Springer, Cham.

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019, October). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 367-372). IEEE.

Kronowitz, S. J. (2007). Immediate versus delayed reconstruction. Clinics in Plastic Surgery, 34(1), 39-50.

Kumar, N. M., & Manjula, R. (2019). Design of multi-layer perceptron for the diagnosis of diabetes mellitus using Keras in deep learning. In Smart Intelligent Computing and Applications (pp. 703-711). Springer, Singapore.

Kumar, S., Joshi, R., & Joge, V. (2013). Do clinical symptoms and signs predict reduced renal function among hospitalized adults?. Ann Med Health Sci Res, 3(3), 492-497.

Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N., & Verma, A. (2020). Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. In Advances in Data Science and Management (pp. 435-442). Springer, Singapore.

Kurian, R. A., & Lakshmi, K. S. (2018). An ensemble classifier for the prediction of heart disease. International Journal of Scientific Research in Computer Science, 3(6), 25-31.

Ladha, L., & Deepa, T. (2011). Feature Selection Methods and Algorithms. International Journal on Computer Science and Engineering (IJCSE), 3(5), 1787 – 1797.

Lakshmanan, B. C., Valarmathi, S., & Ponnuraja, C. (2015). Data mining with decision tree to evaluate the pattern on effectiveness of treatment for pulmonary tuberculosis: a clustering and classification techniques. Sci. Res. J, 3(6), 43-48.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. Machine learning, 59(1-2), 161-205.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lee, S., Schowe, B., Sivakumar, V., & Morik, K. (2012). Feature selection for high-dimensional data with rapidminer. Universitätsbibliothek Dortmund.

Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. The lancet, 379(9811), 165-180.

Levey, A. S., Atkins, R., Coresh, J., Cohen, E. P., Collins, A. J., Eckardt, K. U., ... & Eknoyan, G. (2007). Chronic kidney disease as a global public health problem: approaches and initiatives–a position statement from Kidney Disease Improving Global Outcomes. Kidney international, 72(3), 247-259.

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N., & Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Annals of internal medicine, 130(6), 461-470.

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y., Castro III, A. F., Feldman, H. I., ... & Coresh, J. (2009). A new equation to estimate glomerular filtration rate. Annals of internal medicine, 150(9), 604-612.

Li, H., Luo, M., Luo, J., Zheng, J., Zeng, R., Du, Q., ... & Ouyang, N. (2016). A discriminant analysis prediction model of non-syndromic cleft lip with or without cleft palate based on risk factors. BMC Pregnancy Childbirth, 16(1), 368.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 1-45.

Li, J., Fong, S., Mohammed, S., Fiaidhi, J., Chen, Q., & Tan, Z. (2016). Solving the under-fitting problem for decision tree algorithms by incremental swarm optimization in rare-event healthcare classification. J Med Imaging Health Inform, 6(4), 1102-1110.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 39(12), 11303-11311.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lim, S., Tucker, C. S., & Kumara, S. (2017). An unsupervised machine learning model for discovering latent infectious diseases using social media data. J. Biomed. Inform., 66, 82-94.

Liu, N., Qi, E. S., Xu, M., Gao, B., & Liu, G. Q. (2019). A novel intelligent classification model for breast cancer diagnosis. Information Processing & Management, 56(3), 609-623.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Comput Math Method M..

Liu, Y. Q., Wang, C., & Zhang, L. (2009, June). Decision tree based predictive models for breast cancer survivability on imbalanced data. In 2009 3rd international conference on bioinformatics and biomedical engineering (pp. 1-4). IEEE.

Loh, W. Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23.

Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. Expert Syst. Appl., 42(21), 8221-8231.

Lucasius, C. B., & Kateman, G. (1993). Understanding and using genetic algorithms Part 1. Concepts, properties and context. Chemometrics and intelligent laboratory systems, 19(1), 1-33.

Lukmanto, R. B., & Irwansyah, E. (2015). The early detection of diabetes mellitus (DM) using fuzzy hierarchical model. Procedia Comput. Sci., 59, 312-319.

Luo, T. I. A. N. Y. I., Krishnan, D., & Sen, S. (2018). Parched power: Water demands, risks, and opportunities for India's power sector. World Resources Institute, Washington, DC. Available at: http://www. wri. org/publication/parched-power.

Lysaght, M. J. (2002). Maintenance dialysis population dynamics: current trends and long-term implications. J Am Soc Nephrol, 13(suppl 1), S37-S40.

Mafarja, M., & Sabar, N. R. (2018, June). Rank based binary particle swarm optimisation for feature selection in classification. In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems (pp. 1-6).

Magesh, G., & Swarnalatha, P. (2020). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. Evol Intell, 1-11.

Mahdavi-Mazdeh, M. (2010). Why do we need chronic kidney disease screening and which way to go?. Iran J Kidney Dis, 4(4), 275-281.

Mahdavi-Mazdeh, M., Hatmi, Z. N., & Shahpari-Niri, S. (2012). Does a medical management program for CKD patients postpone renal replacement therapy and mortality?: A 5-year-cohort study. BMC Nephrol, 13(1), 138.

Manikandan, K. (2019). Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization. Pattern Recognit Lett, 125, 432-438.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed, 152, 23-34.

Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Inf. Sci. Syst., 8(1), 7.

Marshland, S. (2009). Machine Learning an Algorithmic Perspective. CRC Press, New Zealand, 6-7.

Mendonça, L. F., Vieira, S. M., & Sousa, J. M. C. (2007). Decision tree search methods in fuzzy modeling and classification. International Journal of Approximate Reasoning, 44(2), 106-123.

Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. Expert Systems with Applications, 72, 335-343.

Mienye, I. D., Sun, Y., & Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Inform. Med. Unlocked, 100307.

Mir, A., & Dhage, S. N. (2018, August). Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthc Inform Res, 22(3), 196-205.

Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Mohebian, M. R., Marateb, H. R., Mansourian, M., Mañanas, M. A., & Mokarian, F. (2017). A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. Computational and structural biotechnology journal, 15, 75-85.

Momete, D. C. (2016). Building a sustainable healthcare model: A cross-country analysis. Sustainability, 8(9), 836.

Most Frequent Cancers in Men and Women. 2008 [sighted 2012 20/01/2012]; Available from: http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900.

Muhammad, G. (2015). Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system. Cluster Comput, 18(2), 795-802.

Nahas, M. E. (2005). The global challenge of chronic kidney disease. Kidney Int., 68(6), 2918-2929.

Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Comput. Sci., 69, 132-142.

Naqvi, G. (2012). A Hybrid Filter-Wrapper Approach for Feature Selection. (Master's Thesis). Orebro University, Sweden.

Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. IEEE Computer Architecture Letters, 26(09), 917-922.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs, Richard D. Irwin, Homewood, Illinois.

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. Artif. Intell. Rev., 33(4), 275-306.

NHANES. Physical Activity and CVD Fitness Data. National Centre for Health Statistics in Centre for Disease Control. https://www.cdc.gov/nchs/tutorials/PhysicalActivity/SurveyOrientation/DataStructure/index.htm. Accessed 27 June, 2020.

NHANES-A. https://wwwn.cdc.gov/Nchs/Nhanes/2009- 2010/DIQ_F.htm. Accessed 27 June, 2020.

NHANES-B. https://wwwn.cdc.gov/Nchs/Nhanes/2013- 2014/DIQ_H.htm. Accessed 27 June, 2020.

Nie, L., Zhao, Y. L., Akbari, M., Shen, J., & Chua, T. S. (2014). Bridging the vocabulary gap between health seekers and healthcare knowledge. IEEE Trans Knowl Data Eng, 27(2), 396-409.

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133-144.

Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L. (2017). Accuracy improvement for diabetes disease classification: a case on a public medical dataset. Fuzzy Inf. Eng., 9(3), 345-357.

Ogunleye, A. A., & Qing-Guo, W. (2019). XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform.

Ordóñez, F. J., de Toledo, P., & Sanchis, A. (2015). Sensor-based Bayesian detection of anomalous living patterns in a home setting. Pers. Ubiquitous. Comput., 19(2), 259-270.

Ottom, M. A., & Alshorman, W. (2019). Heart diseases prediction using accumulated rank features selection technique. J. Eng. Appl. Sci, 14, 2249-2257.

Panda, D., & Dash, S. R. (2020). Predictive System: Comparison of Classification Techniques for Effective Prediction of Heart Disease. In Smart Intelligent Computing and Applications (pp. 203-213). Springer, Singapore.

Panthong, R., & Srivihok, A. (2015). Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Computer Science, 72, 162-169.

Panwar, M., Acharyya, A., Shafik, R. A., & Biswas, D. (2016, December). K-nearest neighbor-based methodology for accurate diagnosis of diabetes mellitus. In 2016 Sixth International Symposium on Embedded Computing and System Design (ISED) (pp. 132-136). IEEE.

Pasadana, I. A., Hartama, D., Zarlis, M., Sianipar, A. S., Munandar, A., Baeha, S., & Alam, A. R. M. (2019, August). Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques. In Journal of Physics: Conference Series (Vol. 1255, No. 1, p. 012024). IOP Publishing.

Paterlini, S., & Minerva, T. (2010, June). Regression model selection using genetic algorithms. In Proceedings of the 11th WSEAS international conference on nural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on Fuzzy systems (pp. 19-27). World Scientific and Engineering Academy and Society (WSEAS).

Pathak, A. K., & Valan, J. A. (2020). A Predictive Model for Heart Disease Diagnosis Using Fuzzy Logic and Decision Tree. In Smart Computing Paradigms: New Progresses and Challenges (pp. 131-140). Springer, Singapore.

Patil, T. R., & Sherekar, S. S. (2013). SantGadgebaba Amravati University Amravati. In International Journal Of Computer Science And Applications (Vol. 6, No. 2).

Paul, A. K., Shill, P. C., Rabin, M. R. I., & Murase, K. (2018). Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. Appl. Intell., 48(7), 1739-1756.

Paulson, R. L., Chang, F. C., & Helmer, S. D. (1994). Kansas surgeons' attitudes toward immediate breast reconstruction: a statewide survey. The American journal of surgery, 168(6), 543-546.

Pei, D., Gong, Y., Kang, H., Zhang, C., & Guo, Q. (2019). Accurate and rapid screening model for potential diabetes mellitus. BMC Medical Inform. Decis. Mak., 19(1), 41.

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., & Johannes, A. (2019). Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. Comput Electron Agr, 161, 280-290.

PID. https://www.kaggle.com/uciml/pima-indians-diabetes- database. Accessed 28 June, 2020.

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J Med Syst, 41(4), 55.

Polat, K., & Sentürk, U. (2018, October). A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. In 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-4). IEEE.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 8, 20991-21002.

Rabby, A. S. A., Mamata, R., Laboni, M. A., & Abujar, S. (2019, July). Machine Learning Applied to Kidney Disease Prediction: Comparison Study. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

Raducanu, B., & Dornaika, F. (2012). A supervised non-linear dimensionality reduction approach for manifold learning. Pattern Recognit, 45(6), 2432-2444.

Raghavendra, U., Fujita, H., Gudigar, A., Shetty, R., Nayak, K., Pai, U., ... & Acharya, U. R. (2018). Automated technique for coronary artery disease characterization and

classification using DD-DTDWT in ultrasound images. Biomedical Signal Processing and Control, 40, 324-334.

Rahman, M. J. U., Sultan, R. I., Mahmud, F., Shawon, A., & Khan, A. (2018, September). Ensemble of Multiple Models for Robust Intelligent Heart Disease Prediction System. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (pp. 58-63). IEEE.

Rajliwall, N. S., Davey, R., & Chetty, G. (2018, December). Machine learning based models for cardiovascular risk prediction. In 2018 International Conference on Machine Learning and Data Engineering (iCMLDE) (pp. 142-148). IEEE.

RamaDevi, G. N., Rani, K. U., & Lavanya, D. (2018, January). Ensemble-based hybrid approach for breast cancer data. In International Conference on Communications and Cyber Physical Engineering 2018 (pp. 713-720). Springer, Singapore.

Ramani, P., Pradhan, N., & Sharma, A. K. (2020). Classification Algorithms to Predict Heart Diseases—A Survey. In Computer Vision and Machine Intelligence in Medical Image Analysis (pp. 65-71). Springer, Singapore.

Rawat, V., & Suryakant, S. (2019). A Classification System for Diabetic Patients with Machine Learning Techniques. Int. J. Math. Eng. Manag. Sci., 4(3), 729-744.

Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 100011.

Reitmaier, T., & Sick, B. (2015). The responsibility weighted Mahalanobis kernel for semi-supervised training of support vector machines for classification. Inf. Sci., 323, 179-198.

Rendell, L., & Seshu, R. (1990). Learning hard concepts through constructive induction: Framework and rationale. Computational Intelligence, 6(4), 247-270.

Riley, G., & Giarratano, J. C. (2005). Expert systems: principles and programming. Thomson Course Technology.

Ripon, S. H. (2019, February). Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

Rubini, L. J. (2015). UCIMachineLearningRepository [http://archive. ics. uci. edu/ml/datasets/Chronic_Kidney_Disease]. Karaikudi. TamilNadu: Algappa University, Department of Computer Science and Engineering.

Rubini, L. J., & Perumal, E. (2020). Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. Int J Imaging Syst Technol.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational psychiatry, 6(10), e921-e921.

Sabahi, F. (2018). Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment. J Biomed Inform, 83, 204-216.

Saha, A., Saha, A., & Mittra, T. (2019, July). Performance Measurements of Machine Learning Approaches for Prediction and Diagnosis of Chronic Kidney Disease (CKD). In Proceedings of the 2019 7th International Conference on Computer and Communications Management (pp. 200-204).

Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access, 6, 29637-29647.

Salekin, A., & Stankovic, J. (2016, October). Detection of chronic kidney disease and selecting important predictive attributes. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 262-270). IEEE.

Sankaranarayanan, R., & Ferlay, J. (2006). Worldwide burden of gynaecological cancer: the size of the problem. Best practice & research Clinical obstetrics & gynaecology, 20(2), 207-225.

Saqlain, S. M., Sher, M., Shah, F. A., Khan, I., Ashraf, M. U., Awais, M., & Ghani, A. (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl Inf Syst, 58(1), 139-167.

Saria, S., Rajani, A. K., Gould, J., Koller, D., & Penn, A. A. (2010). Integration of early physiological responses predicts later illness severity in preterm infants. Sci. Transl. Med., 2(48), 48ra65-48ra65.

Saringat, Z., Mustapha, A., Saedudin, R. R., & Samsudin, N. A. (2019). Comparative analysis of classification algorithms for chronic kidney disease diagnosis. Bulletin of Electrical Engineering and Informatics, 8(4), 1496-1501.

Scheffler, R., Cometto, G., Tulenko, K., Bruckner, T., Liu, J., Keuffel, E. L., ...& Campbell, J. (2016). Health workforce requirements for universal health coverage and the Sustainable Development Goals–Background paper N. 1 to the WHO Global Strategy on Human Resources for Health: Workforce 2030. Human resources for health observer series, (17).

Schreiner, S. J., Imbach, L. L., Werth, E., Poryazova, R., Baumann-Vogel, H., Valko, P. O., ... & Baumann, C. R. (2019). Slow-wave sleep and motor progression in Parkinson disease. Annals of neurology, 85(5), 765-770.

Schwarz, P. E., Li, J., Lindstrom, J., & Tuomilehto, J. (2009). Tools for predicting the risk of type 2 diabetes in daily practice. Horm. Metab. Res., 41(02), 86-97.

Sen, T., & Das, S. (2013). An approach to pancreatic cancer detection using artificial neural network. In Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering-CEEE (pp. 56-60).

Sharaff, A., & Gupta, H. (2019). Extra-tree classifier with metaheuristics approach for email classification. In Advances in Computer Communication and Computational Sciences (pp. 189-197). Springer, Singapore.

Sharma, A., Kulshrestha, S., & Daniel, S. (2017, December). Machine learning approaches for breast cancer diagnosis and prognosis. In 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp) (pp. 1-5). IEEE.

Sharma, M., Bansal, A., Gupta, S., Asija, C., & Deswal, S. (2020). Bio-inspired Algorithms for Diagnosis of Heart Disease. In International Conference on Innovative Computing and Communications (pp. 531-542). Springer, Singapore.

Shi, D., Zurada, J., & Guan, J. (2015, January). A Neuro-fuzzy system with semi-supervised learning for bad debt recovery in the healthcare industry. In 2015 48th Hawaii International Conference on System Sciences (pp. 3115-3124). IEEE.

Shoeb, A. H., & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. In 27th International Conference on Machine Learning (ICML-10) (pp. 975-982).

Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In 2012 Japan-Egypt Conference on Electronics, Communications and Computers (pp. 173-177). IEEE.

Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In Advances in Computing and Intelligent Systems (pp. 195-211). Springer, Singapore.

Shylaja, S., & Muralidharan, R. (2018). Comparative Analysis of Various Classification and Clustering Algorithms for Heart Disease Prediction System. Biometrics and Bioinformatics, 10, 74-77.

Singh, A., & Gupta, G. (2019). ANT_FDCSM: A novel fuzzy rule miner derived from ant colony meta-heuristic for diagnosis of diabetic patients. J Intell Fuzzy Syst, 36(1), 747-760.

Singh, A., & Kumar, R. (2020, February). Heart Disease Prediction Using Machine Learning Algorithms. In 2020 International Conference on Electrical and Electronics Engineering (ICE3) (pp. 452-457). IEEE.

Singh, N. (2005). HPV and Cervical cancer-prospects for prevention through vaccination. Indian J Med Paediatr Oncol, 26(1), 20-23.

Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. Biocybern Biomed Eng, 40(1), 1-22.

Sipes, T., Jiang, S., Moore, K., Li, N., Karimabadi, H., & Barr, J. R. (2014). Anomaly Detection in Healthcare: Detecting Erroneous Treatment Plans in Time Series Radiotherapy Data. Int. J. Semant. Comput., 8(03), 257-278.

Sisodia, D. S., & Verma, A. (2017, November). Prediction performance of individual and ensemble learners for chronic kidney disease. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 1027-1031). IEEE.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Comput. Sci., 132, 1578-1585.

Sobrinho, A., Queiroz, A. C. D. S., Da Silva, L. D., Costa, E. D. B., Pinheiro, M. E., & Perkusich, A. (2020). Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques. IEEE Access, 8, 25407-25419.

Stevens, L. A., & Levey, A. S. (2009). Current status and future perspectives for CKD testing. American Journal of Kidney Diseases, 53(3), S17-S26.

Subasi, A., Alickovic, E., & Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In CMBEBIH 2017 (pp. 589-594). Springer, Singapore.

Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction. In 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-5). IEEE.

Tazin, N., Sabab, S. A., & Chowdhury, M. T. (2016, December). Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. In 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec) (pp. 1-6). IEEE.

Tikariha, P., & Richhariya, P. (2018). Comparative Study of Chronic Kidney Disease Prediction Using Different Classification Techniques. In Proceedings of International Conference on Recent Advancement on Computer and Communication (pp. 195-203). Springer, Singapore.

Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. Jama, 316(5), 533-534.

Trivedi, S. K., & Dey, S. (2014, October). A study of ensemble based evolutionary classifiers for detecting unsolicited emails. In Proceedings of the 2014 conference on research in adaptive and convergent systems (pp. 46-51).

Trivedi, S. K., & Dey, S. (2019). A modified content-based evolutionary approach to identify unsolicited emails. Knowledge and Information Systems, 60(3), 1427-1451.

Tu, M. C., Shin, D., & Shin, D. (2009, October). Effective diagnosis of heart disease through bagging approach. In 2009 2nd International Conference on Biomedical Engineering and Informatics (pp. 1-4). IEEE.

UCI-A. Detrano, R. V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+ disease. Accessed 20 June, 2020.

UCI-B. Janosi, A. Hungarian Institute of Cardiology, Budapest. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+Disease. Accessed 24 June, 2020.

UCI-C. Statlog Heart Dataset. University of California, School of Information and Computer Science. Irvine, CA. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29. Accessed 22 June, 2020.

UCI-D. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart. Accessed 23 June, 2020.

UCI-E. Detrano, R. V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+ disease. Accessed 23 Julne, 2020.

UCI-F. Steinbrunn, W. University Hospital, Zurich, Switzerland. Pfisterer, M. University Hospital, Basel, Switzerland. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease. Accessed 25 June, 2020.

UCIMLR. (1988). https://archive.ics.uci.edu/ml/datasets/ Heart+Disease. Accessed 18 June, 2020.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Med Informat Decis Making, 19(1), 1-16.

Unler, A., Murat, A., & Chinnam, R. B. (2011). mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inf. Sci., 181(20), 4625-4641.

Uyar, K., & Ilhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput. Sci., 120, 588-593.

Vigneswari, D., Kumar, N. K., Raj, V. G., Gugan, A., & Vikash, S. R. (2019, March). Machine Learning Tree Classifiers in Predicting Diabetes Mellitus. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 84-87). IEEE.

Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.

Vijiyarani, S., & Sudha, S. (2013). Disease prediction in data mining technique–a survey. International Journal of Computer Applications & Information Technology, 2(1), 17-21.

Vivekanandan, T., & Iyengar, N. C. S. N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Comput Biol Med, 90, 125-136.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. Pertanika Journal of Science & Technology, 26(1).

Wahba, G., Lin, X., Gao, F., Xiang, D., Klein, R., & Klein, B. E. (1998, September). The Bias-Variance Tradeoff and the Randomized GACV. In NIPS (pp. 620-626).

Wahba, G., Wang, Y., Gu, C., Klein, R., & Klein, B. (1994). Structured machine learning forsoft'classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. Advances in Neural Information Processing Systems, 6, 415-422.

Walser, M. (1994). Assessment of renal function and progression of disease. Curr. Opin. Nephrol. Hypertens., 3(5), 564-567.

Walser, M., Drew, H. H., & Guldan, J. L. (1993). Prediction of glomerular filtration rate from serum creatinine concentration in advanced chronic renal failure. Kidney Int., 44(5), 1145-1148.

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, 267(2), 687-699.

Wang, P., Liu, Z., Gao, R. X., & Guo, Y. (2019). Heterogeneous data-driven hybrid machine learning for tool condition prognosis. CIRP Annals, 68(1), 455-458.

Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86, 105941.

Wang, Y., Chen, S., Xue, H., & Fu, Z. (2015). Semi-supervised classification learning by discrimination-aware manifold regularization. Neurocomputing, 147, 299-306.

Wang, Y., Wu, S., Li, D., Mehrabi, S., & Liu, H. (2016). A Part-Of-Speech term weighting scheme for biomedical information retrieval. J Biomed Inform, 63, 379-389.

Wei, S., Zhao, X., & Miao, C. (2018, February). A comprehensive exploration to the machine learning techniques for diabetes identification. In 2018 IEEE 4th World Forum on Internet of Things (WF-IoT) (pp. 291-295). IEEE.

Weiss, S. M., & Kulikowski, C. A. (1991). Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. Morgan Kaufmann Publishers Inc..

WHO (2007). WHO/ICO Information Centre on HPV and Cervical Cancer (HPV Information Centre). Summary report on HPV and cervical cancer statistics in India 2007. [Last Assessed on 2008 May 1]. Available from: http://www.who.int/hpvcentre.

Wibawa, M. S., Maysanjaya, I. M. D., & Putra, I. M. A. W. (2017, August). Boosted classifier and features selection for enhancing chronic kidney disease diagnose. In 2017 5th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-6). IEEE.

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis, 66(1), 149-153.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in statistics (pp. 196-202). Springer, New York, NY.

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In Soft computing and industry (pp. 25-42). Springer, London.

Wongchaisuwat, P., Klabjan, D., & Jonnalagadda, S. R. (2016). A Semi-Supervised Learning Approach to Enhance Health Care Community–Based Question Answering: A Case Study in Alcoholism. JMIR Med Inf, 4(3), e24.

World Health Organization. (2011). Global status report on noncommunicable diseases 2010.

World Health Organization. (2016). Total expenditure on health as a percentage of gross domestic product (US $). Global Health Observatory.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Inform. Med. Unlocked, 10, 100-107.

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Peer Reviewed: Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev. Chronic Dis., 16, E130.

Xiong, X. L., Zhang, R. X., Bi, Y., Zhou, W. H., Yu, Y., & Zhu, D. L. (2019). Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults. Curr Med Sci, 39(4), 582-588.

Xu, Z., & Wang, Z. (2019, June). A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. In 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI) (pp. 278-283). IEEE.

Yan, Y., Chen, L., & Tjhi, W. C. (2013). Fuzzy semi-supervised co-clustering for text documents. Fuzzy Sets Syst, 215, 74-89.

Yildirim, P. (2015). Filter based feature selection methods for prediction of risks in hepatitis disease. International Journal of Machine Learning and Computing, 5(4), 258.

Yuan, B., & Ma, X. (2012, June). Sampling+ reweighting: boosting the performance of adaboost on imbalanced datasets. In The 2012 international joint conference on neural networks (IJCNN) (pp. 1-6). IEEE.

Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Comput, 22(1), 1-9.

Zelenkov, Y., Fedorova, E., & Chekrizov, D. (2017). Two-step classification method based on genetic algorithm for bankruptcy forecasting. Expert Systems with Applications, 88, 393-401.

Zeynu, S., & Patil, S. (2018). Prediction of Chronic Kidney Disease using Data Mining Feature Selection and Ensemble Method. International Journal of Data Mining in Genomics & Proteomics, 9(1), 1-9.

Zhang, M., Yang, L., Ren, J., Ahlgren, N. A., Fuhrman, J. A., & Sun, F. (2017). Prediction of virus-host infectious association by supervised learning methods. BMC Bioinform., 18(3), 60.

Zhang, Y., Wang, S., Phillips, P., & Ji, G. (2014). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 64, 22-31.

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. Annals of translational medicine, 4(7).

Zriqat, I. A., MousaAltamimi, A., & Azzeh, M. (2016). A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. International Journal of Computer Science and Information Security (IJCSIS), 14(12).

Appendix I

Testimonies of Paper Presentations

Variable Selection in Genetic Algorithm Model with Logistic Regression for Prediction of Progression to Diseases

Publisher: IEEE

Cite This

PDF

Avijit Kumar Chaudhuri ; Anirban Das **All Authors**

Abstract:
Earlier risk assessment and identification of different diseases is the most crucial issue for avoiding and lowering their progression. The researchers typically used the statistical comparative analysis or step-by-step methods of feature selection using regression techniques to estimate the risk factors of diseases. The outcomes from these methods emphasized on individual risk factors separately. A combination of factors, however, is likely to affect the development of disease rather than just anyone alone. Genetic algorithms (GA) can be useful and efficient for searching a combination of factors for the fast diagnosis with best accuracies, especially for a large number of complex and poorly understood factors, as in the case in the prediction of disease development. Our proposed model shows the potential for the application of GA in diagnosing disease and predicting accuracy. Our proposed model demonstrated that the amalgamation of a small subset of input features produces the optimum performance than the use of all the single significant features individually. This model not only predicts the best feature sets and accuracy but also overcome the problem of missing values present in the dataset. Variables more frequently selected by LR might be more important for the prediction of disease development and accuracies by GA.

Published in: 2020 IEEE International Conference for Innovation in Technology (INOCON)

Date of Conference: 6-8 Nov. 2020

Date Added to IEEE *Xplore*: 01 January 2021

ISBN Information:

INSPEC Accession Number: 20284740

DOI: 10.1109/INOCON50539.2020.9298372

Publisher: IEEE

Conference Location: Bangluru, India

International Conference on Computer Science, Engineering and Application (ICCSEA-2020)

Organized by
Department of Computer Science and Engineering,
GIET University, Gunupur

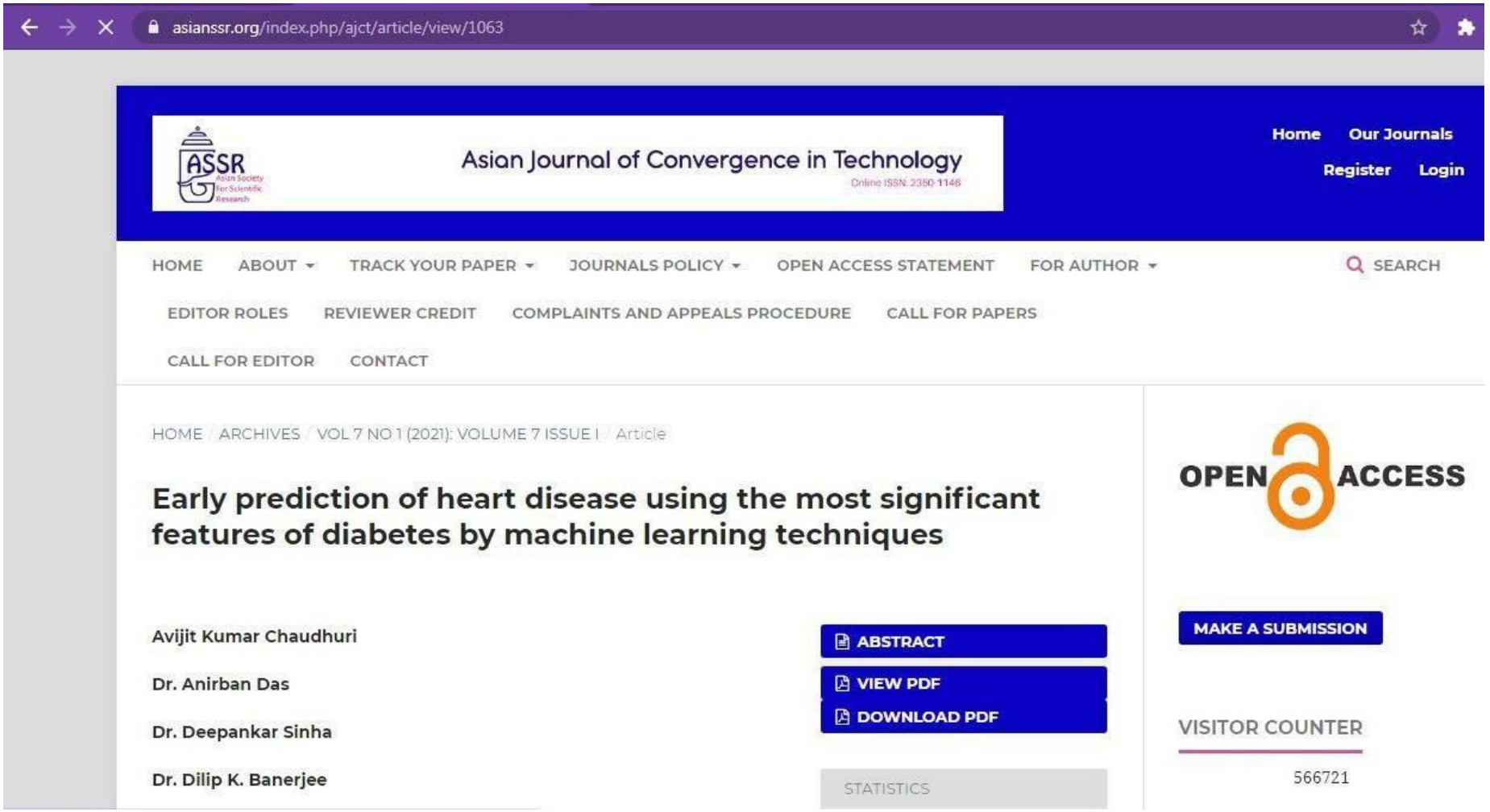
Certificate

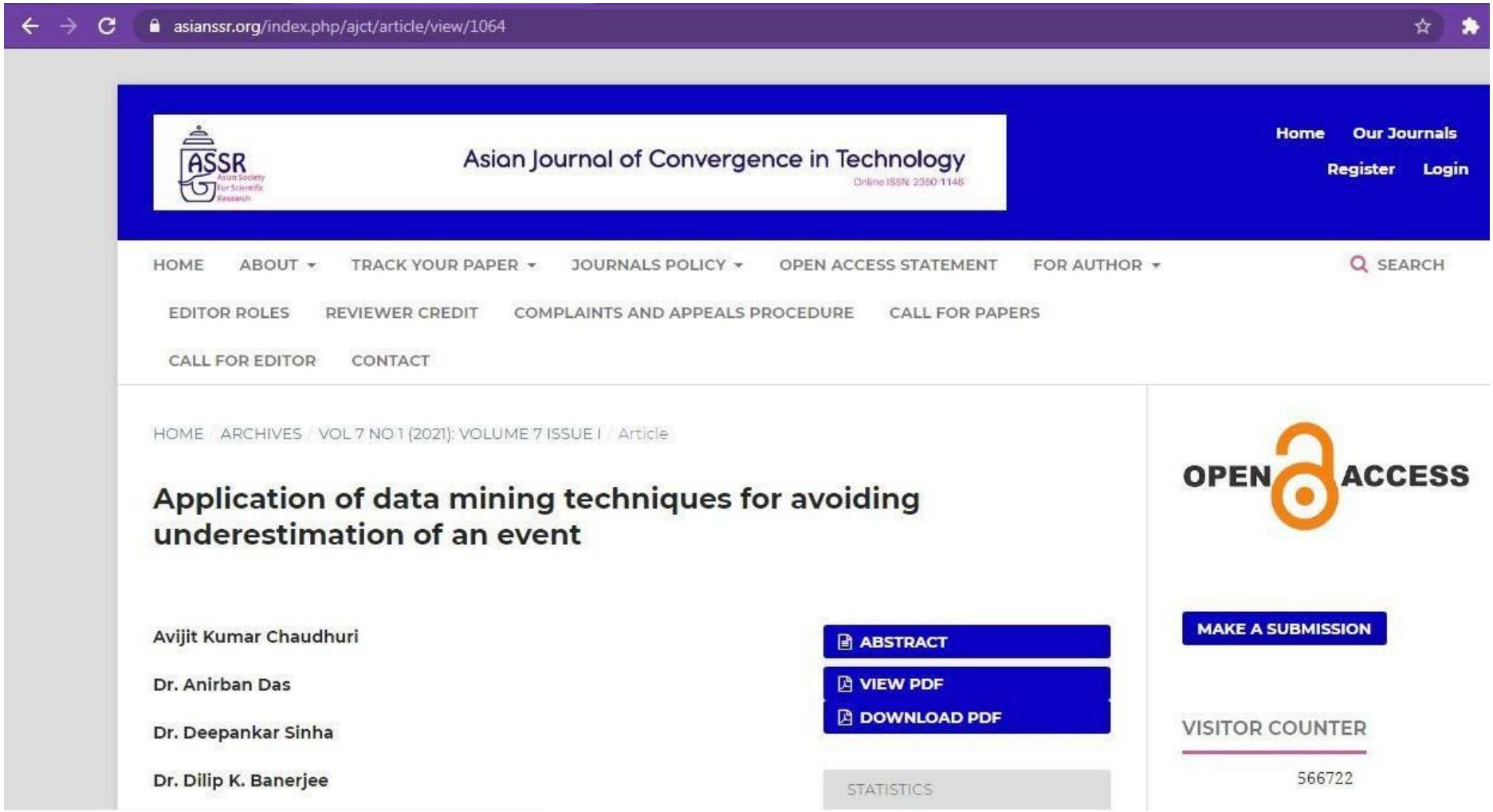
This is to certify that Mr. /Ms. /Prof. /Dr. **Madhumita Addy, Avijit Kumar Chaudhuri and Anirban Das** affiliated to **Imerit Technology Pvt. Ltd., Techno Engg. College Banipur and University of Engg. & Mang. Kolkata India** has participated in the "International Conference on Computer Science, Engineering and Application" held at GIET University, Gunupur, India on 13th & 14th March 2020. He / She also **presented a Paper** / Delivered a Keynote Talk / Co-ordinated / Organised / Chaired a technical session titled **"Role of Data Mining techniques and MCDM model in detection and severity monitoring to serve as precautionary methodologies against 'Dengue'".**

Conference Chair

Appendix II

Testimonies of Paper Publications


asianssr.org/index.php/ajct/article/view/1063
ASSR
Asian Society For Scientific Research
Asian Journal of Convergence in Technology
Online ISSN: 2350-1146
Home
Our Journals
Register
Login
HOME
ABOUT
TRACK YOUR PAPER
JOURNALS POLICY
OPEN ACCESS STATEMENT
FOR AUTHOR
SEARCH
EDITOR ROLES
REVIEWER CREDIT
COMPLAINTS AND APPEALS PROCEDURE
CALL FOR PAPERS
CALL FOR EDITOR
CONTACT
HOME / ARCHIVES / VOL 7 NO 1 (2021): VOLUME 7 ISSUE I / Article
Early prediction of heart disease using the most significant features of diabetes by machine learning techniques
Avijit Kumar Chaudhuri
Dr. Anirban Das
Dr. Deepankar Sinha
Dr. Dilip K. Banerjee
ABSTRACT
VIEW PDF
DOWNLOAD PDF
STATISTICS
OPEN ACCESS
MAKE A SUBMISSION
VISITOR COUNTER
566721


asianssr.org/index.php/ajct/article/view/1064
ASSR
Asian Society For Scientific Research
Asian Journal of Convergence in Technology
Online ISSN: 2350-1146
Home
Our Journals
Register
Login
HOME
ABOUT
TRACK YOUR PAPER
JOURNALS POLICY
OPEN ACCESS STATEMENT
FOR AUTHOR
SEARCH
EDITOR ROLES
REVIEWER CREDIT
COMPLAINTS AND APPEALS PROCEDURE
CALL FOR PAPERS
CALL FOR EDITOR
CONTACT
HOME / ARCHIVES / VOL 7 NO 1 (2021): VOLUME 7 ISSUE I / Article
Application of data mining techniques for avoiding underestimation of an event
Avijit Kumar Chaudhuri
Dr. Anirban Das
Dr. Deepankar Sinha
Dr. Dilip K. Banerjee
ABSTRACT
VIEW PDF
DOWNLOAD PDF
STATISTICS
OPEN ACCESS
MAKE A SUBMISSION
VISITOR COUNTER
566722

Original Article | Published: 11 April 2021

A novel enhanced decision tree model for detecting chronic kidney disease

Avijit Kumar Chaudhuri, Deepankar Sinha, Dilip K. Banerjee & Anirban Das

Network Modeling Analysis in Health Informatics and Bioinformatics **10**, Article number: 29 (2021) | Cite this article

Abstract

Prediction of diseases is sensitive as any error can result in the wrong person's treatment or not treating the right patient. Besides, some features distinguish a disease from curable to fatal or curable to chronic disease. Data mining techniques have been widely used in health-related research. The researchers, so far, could attain around 97 percent accuracy using several methods. Some researchers have demonstrated that the selection of correct features increases the prediction accuracy. This research work propose a method to distinguish between chronic and non-chronic kidney disease, identify its crucial features without reducing the accuracy of prediction, and a prediction algorithm to eliminate the possibility of under or overfitting. This study uses the recursive feature elimination (RFE) method that selects an optimal subset of

Outline

ELSEVIER

Machine Learning with Applications

Volume 3, 15 March 2021, 100011

Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development

Arkadip Ray [a], Avijit Kumar Chaudhuri [b]

[a] Department of Information Technology (IT), Government College of Engineering & Ceramic Technology, West Bengal, India

[b] Department of Computer Science and Engineering, Techno Engineering College Banipur, Habra, West Bengal, India

Received 4 July 2020, Revised 18 November 2020, Accepted 18 November 2020, Available online 17 December 2020.

Gmail Avijit Chaudhuri <c.avijit@gmail.com>

AJCT Acceptance Notification

5 messages

INOCON 2020 <inocon2020@easychair.org> Thu, Nov 19, 2020 at 8:41 PM
To: Avijit Kumar Chaudhuri <c.avijit@gmail.com>

Dear Avijit Kumar Chaudhuri,

Paper ID : 875

Happy to share with you that the paper entitled " Early Detection of Cardiovascular Disease in Patients with Chronic Kidney Disease using Data Mining Techniques " is accepted in -Asian Journal of Convergence of Technology (AJCT) with no publication cost.

Kindly mail your the source paper(Doc/Docx file) by 21/11/2020

Format Link : https://drive.google.com/file/d/1U34GCag8DHRygmZk95CQsl8VpesM1giO/view?usp=sharing

Note :
1. Its should be in same format in sent above.
2. File must mailed to inoconf@gmail.com with doc/docx file only , its not must be in protected mode must able to edit if anything by editor.
3. The papers will be published in Volume 3 Issue 3 edition in December first week.
4. While mailing to above given mail id kindly rename your file with paper ID (ex: 434_AJCT)

SPL note : Earlier uploaded word files is in protected mode which uploded in google form.

Thanks & Regards
Publication Chair AJCT
+91-8999689262

Avijit Chaudhuri <c.avijit@gmail.com> Thu, Nov 19, 2020 at 8:53 PM
To: Arkadip Ray <arka1dip2ray3@gmail.com>
Cc: "Dr. Anirban Das" <anirban-das@live.com>, dkbanrg@gmail.com

[Quoted text hidden]

Avijit Chaudhuri <c.avijit@gmail.com> Thu, Nov 19, 2020 at 9:15 PM
To: INOCON 2020 <inocon2020@easychair.org>
Cc: Arkadip Ray <arka1dip2ray3@gmail.com>, "Dr. Anirban Das" <anirban-das@live.com>, dkbanrg@gmail.com, INNOCON CONFERENCE <inoconf@gmail.com>

Mam/Sir,
Thanking you for considering our paper(875_AJCT) for publication in your esteemed journal. One copy of the article in .doc format is attached herewith this mail for your kind reference.
Thanking you,
With kind regards
Avijit Kumar Chaudhuri
[Quoted text hidden]

875_AJCT.doc
537K

International Journal of Recent Technology and Engineering
ISSN: 2277-3878 (Online) | Exploring Innovation | A Key for Dedicated Services
Published by Blue Eyes Intelligence Engineering & Sciences Publication

G:18-19-20, Block-B, Tirupati Abhinav Homes, Damkheda, Bhopal (Madhya Pradesh)-462037, India
Website: www.ijrte.org Email: submit2@ijrte.org
+91-9109122902 | +91-9109122902 | +91-9109122902

CERTIFICATE

This certifies that the research paper entitled '**An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research'** authored by **'Avijit Kumar Chaudhuri, Deepankar Sinha, Kousik Bhattacharya, Anirban Das'** was reviewed by experts in this research area and accepted by the board of 'Blue Eyes Intelligence Engineering and Sciences Publication' which has published in **'International Journal of Recent Technology and Engineering (IJRTE)'**, ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020. Page No.: 1096-1106.

The Value of Citation (VoC) IJRTE is 6.04 for the year 2019. Your published paper and Souvenir are available at: https://www.ijrte.org/download/volume-8-issue-6/

Jitendra Kumar Sen
(Manager)

Blue Eyes Intelligence Engineering and Sciences Publication
BHOPAL

Dr. Shiv Kumar
(CEO)

Gmail Avijit Chaudhuri <c.avijit@gmail.com>

Paper ACCEPTANCE and AUTHOR REGISTRATION@AMLTA2020 || Manipal University Jaipur || 13-15 February 2020

9 messages

Advanced Machine Learning Technologies [MU - Jaipur] <amlta2020@jaipur.manipal.edu>
To: "c.avijit@gmail.com" <c.avijit@gmail.com>

Wed, Oct 30, 2019 at 12:02 PM

Dear AVIJIT CHAUDHURI,

Greetings and Congratulations!

Your paper having **Paper ID: 111** titled '**Identifying the association rule to determine the possibilities of Cardio Vascular Diseases(CVD)'** has been selected for Oral presentation at AMLTA2020 to be held at Manipal University Jaipur from 13-15 February 2020. The reviewers comments are as attached for your ready reference. Please complete all the formalities of revising your paper as per the reviewers comments and register for the conference by **November, 15, 2019** (early bird registration deadline) so as to enable us to meet the **subsequent deadlines from the Springer publishing**.

General comments and considerations:

1- Plagiarism/similarity report should be less than 20%

2- Provide high quality Figures

3- Be sure that all references are complete and cited within the text

4- Add 3-5 keywords (after the abstract)

5- Follow the Springer template

6- 10 pages max length.

Please make use of the attached CR form and following are some of the important guidelines reproduced again for your ready reference.

1. Revise the paper based on the reviews comments and general comments. Then Sign the attached copyright form by "**Hand**" and update the scanned copy in the final folder.

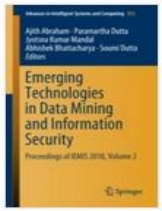
2. Complete the Registration process by paying the registration fee. Visit the following URL for detail:

http://amlta.com/2020/registration/

3. Once registration fee is paid and registration is complete, please fill the following form for us to verify at our end (attach the zipped final folder as mentioned in the step 4 below):

https://tinyurl.com/yxwxa5e8

4. Create a Folder (AMLTA2020 – your Paper ID) which must contain the following:



Emerging Technologies in Data Mining and Information Security pp 519-532 | Cite as

Identification of the Recurrence of Breast Cancer by Discriminant Analysis

Authors | Authors and affiliations

Avijit Kumar Chaudhuri, D. Sinha, K. S. Thyagaraj

Conference paper
First Online: 02 September 2018

1 Citations 524 Downloads

Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 813)

Abstract

Breast cancers enact one of the deadliest diseases that make a high number of deaths every year. It is a special type of all cancers and the primary reason for women's deaths globally (Bangal et al. Breast, carcinoma in women—a rising threat [1]). In the medical field, data mining methods are widely used in diagnosis and analysis to make decisions exclusively. Given the relationship between the degree of malignancy and recurrence of breast cancer, and given the typical model of breast cancer spread, it should be the principal goal of early detection by which cancers can be identified when they are small and node-negative. Most of the researches in the field of breast cancer have focused on predicting and analyzing the disease. There is not much evidence in study of recurrence of the disease. This paper aims at developing an approach to predict and identify the probability of recurrence of breast cancer for the patients with greater accuracy.

Appendix III

Plagiarism Report

grammarly Report: thesis Avijit4

thesis Avijit4

by Anupam Mukherjee

General metrics

52,143	7,930	1007	31 min 43 sec	1 hr 1 min
characters	words	sentences	reading time	speaking time

Score

91

This text scores better than 91% of all texts checked by Grammarly

Writing Issues

295	99	196
Issues left	Critical	Advanced

Plagiarism

3 %

25 sources

3% of your text matches 25 sources on the web or in archives of academic publications