Journal of Applied Science and Education, 2023,
Vol. 03, Iss. 02, S. No. 004, pp. 1-9
ISSN (Online): 2583-1372

Heart Disease Prediction by using Random Forest Classifier

Anamta Siddiqui[1], Syed Wajahat Abbas Rizvi[2]

[1,2]Amity School of Engineering & Technology Lucknow, Amity University Uttar Pradesh, India
[1]anamtasiddiqui29@gmail.com, [2]swarizvi@lko.amity.edu

How to cite this paper: A. Siddiqui and S. W. A. Rizvi, "Heart Disease Prediction by using Random Forest Classifier," *Journal of Applied Science and Education (JASE)*, Vol. 03, Iss. 02, S. No. 004, pp 1-9, 2023.

https://doi.org/10.54060/jase.v3i2.29

Received: 19/03/2023
Accepted: 21/05/2023
Published: 25/11/2023

Abstract

This research presents data on a Machine Learning-based Artificial Intelligence system used in predicting cardiac illness. In this research, we learn how advances in machine learning have improved our ability to foresee who will and will not get heart disease. In both developed and less developed, non-industrialized countries, cardiovascular dis- eases are majorly the main reason of immortality for decades. Reducing mortality from cardiovascular infections requires both early detection and constant clinical supervision. However, it is unrealistic to expect accurate, consistent patient screening, and having a specialist confer with a patient for 24 hours isn't feasible due to the additional knowledge, time, and training it would require. Here, we have used ML algorithms and methods which are likely as linear regression, Random Forest, Decision tree, SVM, KNN, and others to construct and explore models for coronary sickness expectancy via the various cardiac attributes of patient and to spot impending coronary ill-ness. For more accurate diagnosis of heart infections, a Random Forest is developed. Due to its near-perfect accuracy in data preparation, this application necessitates thorough data analysis.

Keywords

Heart disease prediction, Machine learning, Linear regression, Random Forest Classifier Algorithm.

1. Introduction

Coronary heart disease, which can include heart attacks, is responsible for the deaths of a shockingly high number of people each year. This might take place as a consequence of the weakening of the cardiac muscle. In a similar vein, the collapse of cardio- vascular function might be interpreted as the disappointment of the heart in its ability to siphon blood. CAD is an abbreviation that stands for coronary artery disease, which is another name for coronary ailment [2]. Coronary artery disease (CAD) is a condition that can develop when the blood supply to the coronary arteries is compromised. A diagnosis of heart disease can be made based on the presence of symptoms which are pain in chest, issue of high bp, irregular heartrates, heart

failure, and many others [1]. There are numerous varieties of heart diseases, each of which might have a unique set of symptoms. Like:

i. “Coronary illness in veins: chest torment, windedness, torment in neck throat.”

ii. “Heart sickness brought about by unusual pulses, slow heartbeat, distress, chest torment, and so on.”

The majority of people experience chest pain, windedness, inconvenient side effects, chest pain, and so on as usual side effects [26]. The most common adverse effects are discomfort in the chest, feeling winded, and passing out. Coronary artery disease can be caused by a number of factors, including genetic predispositions, high blood pressure, diabetes, smoking, drug use, and alcohol consumption [5, 8]. Symptoms including fever, weakness, a dry hack, and skin rashes can indicate an internal layer disease in patients with coronary disease. Microorganisms, infections, and parasites are the three main causes of cardiovascular disease. Kinds of coronary illness: Heart failure, Hypertension, Coronary supply route sickness, Cardiovascular breakdown, Heart contamination, Intrinsic coronary illness, less intensity of heartbeat, Stroke happening coronary illness, angina pectoris [4, 29].

Presently in one day there are so many mechanized strategies that identify coronary illness as information mining, AI, profound learning, and so forth [22]. In keeping with these ideas, the next section of this paper will provide a concise discussion regarding AI techniques. During this process, we train the datasets by making use of the AI stores. There are a few risks related criteria’s that can help assure whether it either or is neither that someone get heart disease [3]. “Age, sex, blood pressure, cholesterol level, family history of cardiovascular disease, diabetes, smoking, alcohol use, being overweight, heart rate, and chest pain” are all risk factors for coronary heart illness [30].

Python has been utilized in reference to do the duty in determining the presence of cardiac diseases. The dataset consisted of a few different components, including cholesterol, treetops, sex, age, & a few more [38]. During the course of the project, a no. of other importation in libraries, including “matplotlib, NumPy, Pandas, alerts, and a great number of others,” were utilized [6]. The outcomes of the predefined dataset evaluation were examined with the use of python-based tools like correlation matrices, histograms, SVMs, K-Nearest Neighbor’s Classifiers, Random-Forest Classifiers, and Decision Tree Classifiers. Python is widely known as being an open-source language that facilitates the formation of novel solutions for the healthcare industry, which in turn yields better outcomes for patients and ultimately improves the quality of care provided to them [9].

Notwithstanding, the language likewise conforms to the HIPAA agenda for guaranteeing the wellbeing of clinical data [26]. The significant reasons for coronary illness are diabetes, stoutness, unfortunate eating routine, overweight, unnecessary liquor use, and actual inertia [31]. Therefore, coronary illness includes arrhythmia, which is considered to be atherosclerosis, which is the tightening through the conduits brought about by an irregular heart cadence [10]. Certain people will experience these side effects whenever they have an episode of coronary heart disease. In addition, pain which travel through arm, dazedness & unsteadiness, a sore throat, wheezing, an excessive sweat formation are all potential side effects [24]. "Heart Disease, strokes, and coronary disease, sometimes called cardiovascular breakdown and coronary supply route illness," are far more common in those over the age of 65 than they are in younger people [7].

Heart Disease is a kind of disease where symptoms are visible but can’t be easily understood by the person. Sometimes it’s too late to detect the symptoms and root cause of the disease which leads to the death of the patient whereas if these symptoms are detected in the initial state, then the person dying from cardiovascular disease can see cured and saved. The need of having a heart disease prediction system is to have a person analyzed on the basis of several cardiovascular attributes. The automatic prediction using machine learning makes it more accurate to diagnose the disease related to heart which in turn makes the treatment more effective.

This paper is organized in section wise format where section 1 leads to the introductory part of the topic and need of hav-

ing this research paper. Section 2 describes the actual problem statement of the topic followed by the objective of having the heart disease prediction. The fundamental target of this report was to decide critical threatening factors in view of clinical dataset which might prompt heart illness. Section 3 presents the literature review which highlights the analysis work of different clinical and cardiovascular diseases. The UCI AI Data set was used where are 303 rows and 14 columns were used for prediction along with 6 quantitative qualities [21]. Section 4 is all about the methodology approach of this report with the overview analysis. It is used to validate reliability framework; its analysis and prediction is showcased in section 5 of this report. Finally concluded the work done in section 6.

2. Problem Statement

The process fir the checking of heart disease is the more important aspect of the circumstance. Despite the availability of coronary disease prediction tools, some of them may be too expensive or ineffective for routine use in humans [34]. Finding out about cardiac conditions sooner rather than later helps reduce the likelihood of death and other associated symptoms [11]. However, it is ludicrous to expect to screen patients consistently and accurately in all cases, and it is not possible to have a specialist confer with a patient for 24 hours due to the additional knowledge, time, and skill it would require [27]. Since we have a lot of information in this day and age, we can utilize different AI calculations to break down the information for buried designs. Secret examples can be utilized for wellbeing finding in restorative information [25, 32].

One of the main purposes is that the python system can assist with seeming to be OK and support computational offices in extricating important experiences from the data over the medical services areas [12]. Python is viewed as one of the most famous programming dialects for creating medical services applications Elevated degrees of LDL cholesterol, or "terrible" cholesterol, can leads in widely recognized type of coronary illness, coronary course sickness (computer aided design). Patients can encounter side effects, for example, chest torment, windedness, and weariness [13]. Therefore, early detection of cardiovascular diseases can help in pursuing a new way of living in high-risk patients, thereby reducing the confounds, which is a remarkable achievement in the medical field [28].

2.1. Objective

The fundamental target of developing this report is:

a. To foster ML model to foresee future chance of heart illness by executing Strategic Relapse.
b. To decide critical threatening factors in view of clinical dataset which might prompt heart illness [14].
c. To investigate include determination strategies and grasp their functioning guideline. are the primary causes of heart disease.

3. Literature survey

While onset in developing improvement in the area of clinical aspect close by ML different examinations and explores have been completed in these new year's delivering the important critical papers. The paper proposes coronary illness expectation utilizing "K Star, J48, SMO, and Bayes Net and Multi-facet perceptron utilizing WEKA programming." In light of execution from various element "SMO (89% of exactness) and Bayes Net (87% of precision)" accomplish ideal execution than K Star, Multi-facet perceptron and J48 methods utilizing k-crease cross approval [15, 21]. The exact execution accomplished by those calculations are yet not palatable [36]. So that assuming the presentation of exactness is worked on more to give player choice to conclusion sickness in an examination directed utilizing Cleveland dataset for heart infections which has around 303 examples with a utilized 10-overlap Cross Approval, taking into account 13 credits, executing 4 unique calculations, they finished up Naive Bayes and Random Forest leads to most extreme exactness of 95 percent [18]. "Using the comparable dataset

of Framingham, Massachusetts, the examinations were done utilizing 4 models and were prepared and tried with greatest precision K Neighbors Classifier: 87%, support Vector Classifier: 83%, decision Tree Classifier: 79%."

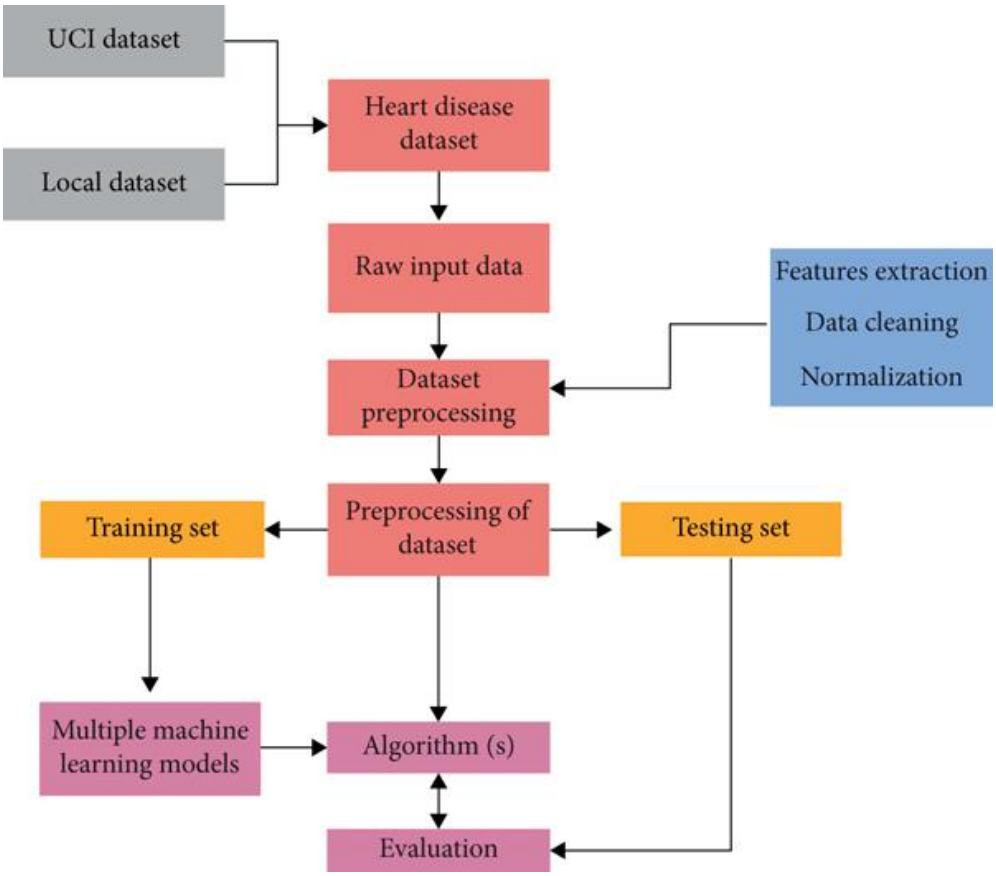

Figure 1. Detecting heart disease

Gomathi involved Naïve Bayes and Decision tree information digging procedures for anticipating various sorts of sicknesses. They primarily focused on forecast of heart illnesses, diabetics, and bosom disease [16]. The outcomes were gotten from the disarray measurements. "Miranda et al. proposed Naïve Bayes classifier approach for the expectation of the cardiovascular sick- nesses." The creators have considered not many significant gamble factors for choosing the coronary illness. Avinash Golande investigates the various ML methods that can be applied to the ranking of cardiovascular disease [19]. Studies were conducted thinking about the accuracy and precision of the decision Tree, linear regression, and K-NN which in turn utilized for sorting. Upon on the search of this study, it can be told that Decision Tree achieves the highest levels of accuracy when using a combi- nation of different processes and boundary adjustments [17]. A ML model comparing five different computations has been planned by Fahd Saleh Alotaibi [35]. The Quick Digger tool was used, and the results were more precise than those obtained using MATLAB or the Weka tool. Classification algorithms such as SVM, RF, NB, and DT were evaluated, along with some lesser-known ones [20]. The most precision which comes of the decision tree algorithm was most great. Many people have worked on disease prognosis frameworks employing various AI calculations for clinical purposes.

3.1. Data Set

During the course of the investigations, the heart dataset that is housed in the UCI AI vault was used. There are 303 rows and 14 columns in the data set. There are 6 quantitative qualities and 8 absolute credits. The table provides a visual representation of the dataset. Selective patients in this dataset range in age from 29 to 79. Patients who are male are assigned an orientation esteem of 1, and patients who are female are assigned an orientation esteem of 0. There are four distinct types of chest pain that might be interpreted as indicators of coronary artery disease. The heart muscle of people with type 1 angina suffers from a lack of oxygen because to blocked arteries of heart. Angina type 1 is characterized by chest issue that occurs in response to intense mental or physical stress. The chest pain that is not due to angina may have other causes and may not always be a sign of coronary disease. The fourth type, which is known as asymptomatic, might not even be a consequence of coronary diseases. The following term. The examination of the pulse will be represented by the notation trestbps. The cholesterol count or Chol. Values of 1 and 0 are assigned to fasting glucose levels of 120 mg/dl and higher, respectively. Fbps stands for fasting glucose. Resting electrocardiogram result (restecg), highest heart rate (thalach), activity-induced angina (1 for pain, 0 for no pain), exercise-induced ST de- pression (old peak), exercise-induced ST depression (slant), nos. of large vessels visualized by fluoroscopy (ca), duration of (thal) in min, & cardiac output (num) are all recorded. For healthy people, the class property is 0, but for those who have been diagnosed with coronary disease, it is 1.

4. Proposed Methodology

In this section, we detailed the procedures that were followed during the experimentation phase of this study. We have mentioned how we have examined the major risk factors for the experiment, as well as the approaches that we have utilized for the prediction of the cardiac illness.

4.1. Libraries importation and data sets loading

This method leads to libraries importing like as NumPy and pandas & afterward stacks the dataset. The most used data form is csv by and large involved design for that AI information is thought of as introduced. This CSV record is utilized for consequently allocating names. In any case, it is marked in the event that the document doesn't have an initial, every segment appear in dataset physically means the characteristics.

4.2. Analysis of data used

EDA is information analyzed utilized in order for acquire a best knowledge on the information & search for the information. That's like a narrative, but for analysts. It takes into account the disclosure of patterns and impressions contained within the material through the use of visual approaches. In addition to this, EDA is sometimes utilized as the initial step towards quite a lengthy demonstration process. It will do an investigation into the dataset as well as carry out an exploratory information examination. It is followed by taking care of missing value, anomaly treatment, encoding all out factors for normal, finally making and deleting copies.

4.3. Evaluation of data

Python considered as a numerical information processing language mostly used Pandas Information Edge sets in order to store information. Python is introduced when data is evaluated, and it is this language that is introduced. The job involved with information inspection includes bringing in, washing, and changing over material in preparation for auditing.

4.4. Data Cleaning

Data cleansing is known as planning information for examination by removing the crude information that set up the information envisioning information & anticipating information. It is the technique for eliminating misleading, tainted, inappropriately organized, and repetitive information from a dataset that will somehow show inaccurate, ruined, and deficient.

4.5. Correlation Matrix-Plot

The strength of a direct association can be depicted using a covariance framework, which is measured using a concept called the relationship. The strength a direct link between two variables, including its direction can be summarized using the Connection Matrix, which accepts values between 1 and +1. The connection lattice's component that displays the relationship between the coefficients is called a component. It is possible to think of a particular irregular variable as being correlated with each and every one of its properties. By visualizing the connection structure as an intensity map, this presents an excellent tool for in-depth investigation into the connections that exist between the various highlights.

A graph is used to express the relationship that exists between factors such as "age, sex, cp, trestbps, Chol, FBS, restecg, thalac, exang, old peak, slope, and ca." The direct relation that exists between 2 continuous variables can be characterized with the help of the correlation that exists inside the dataset.

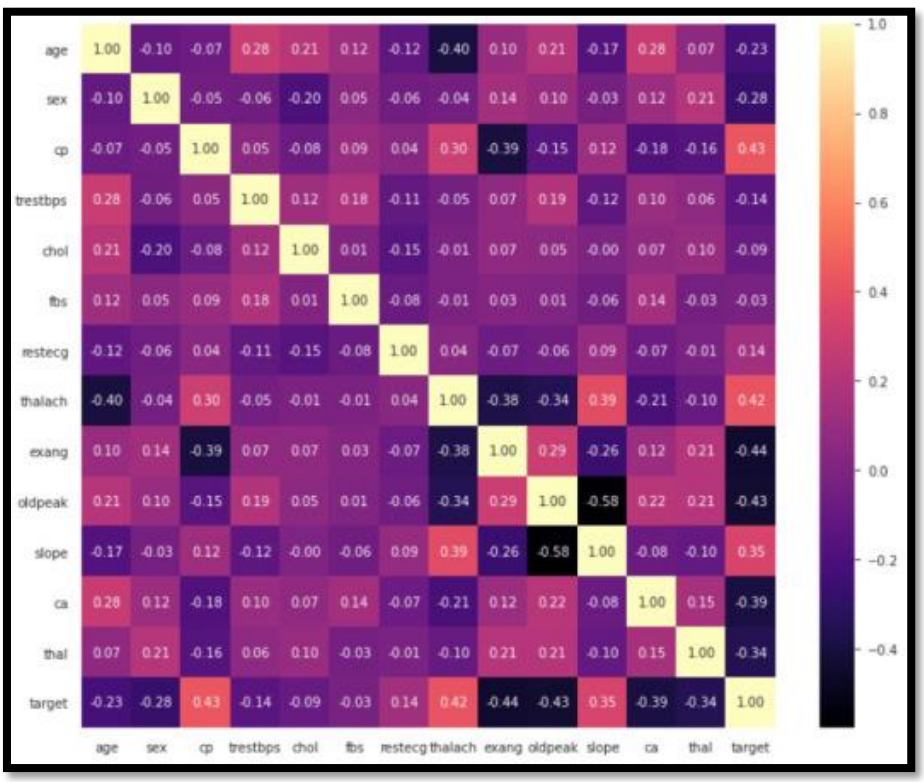

Figure 2. Correlation Matrix Plot

5. Results and Discussion

The purpose behind this study means to compare the accuracy of different ML algorithms in order to identify the highly reliable method of determining whether a given individual would develop coronary disease. processed using the UCI dataset

using “Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest.” Python was used to divide the dataset in data training and testing, build the models, & record any discrepancies. Underneath, we see a relationship between how the calculations are shown and their precision scores, which are introduced in a table [18]. Models were trained and their accuracy was measured utilizing Python after the dataset was partitioned into training and test sets. A comparison of the two algorithms' performance is shown below, along with their respective accuracy scores, that has shown in the table.

Table 1. Table of Accuracy of Different Algorithms

Algorithm	Accuracy
Logistic Regression	85.25 %
Naïve Bayes	85.5 %
Support Vector Machine	81.7%
K-Nearest Neighbor	67.21%
Decision Tree	81.7%
Random Forest	95.8 %

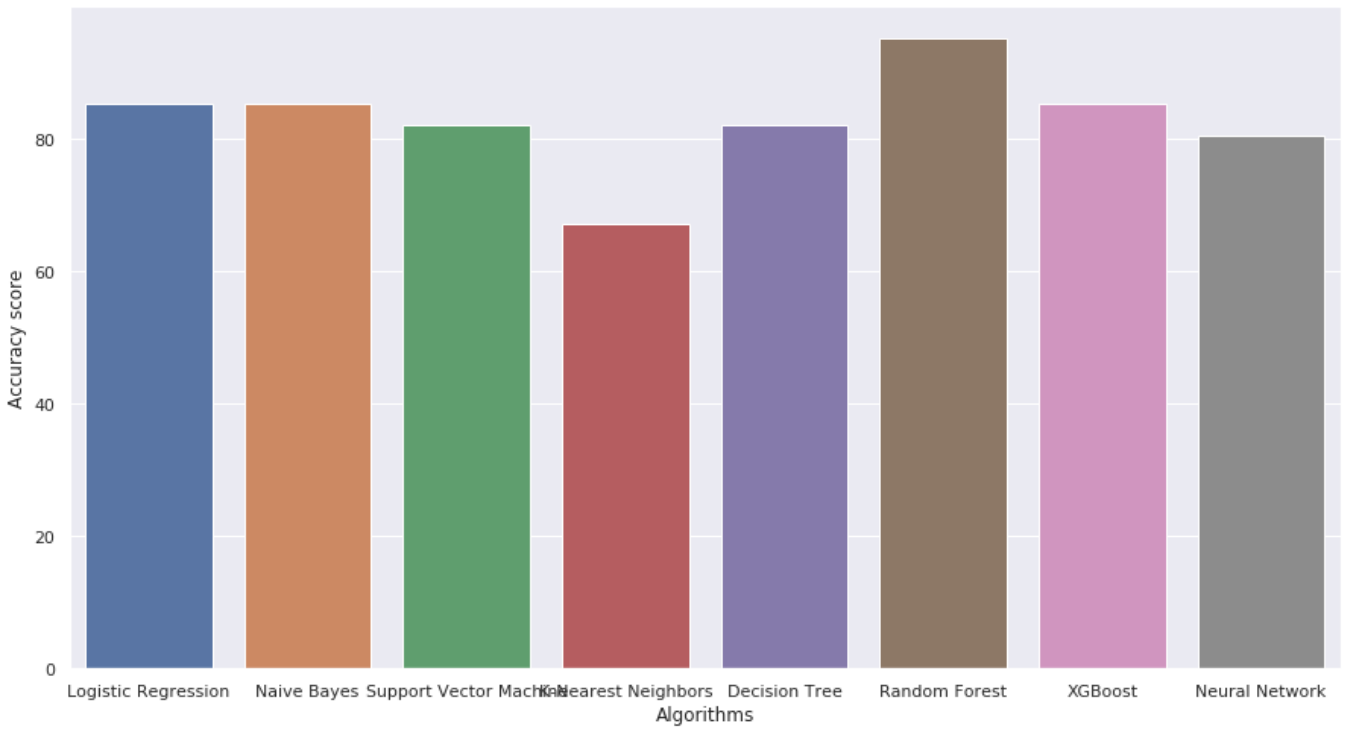

Figure 3. Random forest has more accuracy

6. Conclusion

The purpose of this study is to compare the accuracy of diff. ML algorithms in order to identify the highly reliable method of determining whether a given individual would develop coronary disease. processed using the UCI dataset using “Logistic Re-

gression, Naive Bayes, Support Vector Machine, K-Nearest neighbor, Decision Tree, and Random Forest. Python was used to divide the dataset into training and testing data, build the models, and record any discrepancies." Underneath, we see a relation- ship between how the calculations are shown and their precision scores, which are introduced in a table. Models were trained and their accuracy was measured utilizing Python after the dataset was partitioned into training and test sets. A comparison of the two algorithms' performance is shown below, along with their respective accuracy scores, which are shown in the table.

References

[1]. T. Das and -. Sengur, "Effective analysis of coronary illness through AI models," in *Master frameworks with appli- cations*, 2009.

[2]. J. Vanisree, "Choice Help model for Coronary illness anticipation in light of early indications of 8-51 patients utilizing parallel grouping," in *Global Diary of PC Applications*, 2011.

[3]. Zhang- "Concentrates on utilization of Help Vector Machines in coronary illness expectation model", Electromagnetic Field Issues and Applications, 6th Worldwide Meeting (ICEF), IEEE 2012.

[4]. H. Elshazly, "Lymph sicknesses expectation in view of help vector machine calculation," in *PC Designing and 24 IITM Diary of the executives and IT Frameworks ninth Global Meeting (ICCES)*, 2014.

[5]. B. Kumar and Y. Paul- *"Clinical Utilizations of AI Calculations*. UIET, Kurukshetra College, 2016.

[6]. R. Symbol and V. Kumar- *"Profound Learning in medical services*. UIET, Kurukshetra College, 2018.

[7]. L. Loku, B. Fetaji, A. Krstev, M. Fetaji, and Z. Zdravev, "Using python programming for assessing and solving health management issues"," *Southeast Eur. J. Sustain. Dev*, vol. 4, no. 1, 2020.

[8]. P. Mathur, "Overview of machine learning in healthcare"," in *Machine Learning Applications using Python*, Berkeley, CA: A Press, pp.1–11, 2019.

[9]. P. Guleria and M. Sood, *Intelligent learning analytics in healthcare sector using machine learning", Machine Learning with Health Care Perspective*. Cham: Springer, 2020.

[10]. B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492-499, 2017.

[11]. J. Mcpadden, T.J. Durant, D. R. Bunch *et al.*, "Health care and precision medicine research: analysis of a scalable data science" plat- form," *J. Med. Internet Res*, vol. 21, no. 4, pp. e13043–e13045, 2019.

[12]. T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare"," *Future Healthcare J*, vol. 6, no. 2, p. 94, 2019.

[13]. H. Mayfield, C. Smith, M. Gallagher, and M. Hockings, "Use of freely available datasets and machine learning methods in predicting deforestation"," *Environ. Model. Softw*, vol. 87, pp. 17–19, 2017.

[14]. J. Zaidi, "Predicting heart disease with classification machine learning algorithms"," *Towards Data Sci*, 2020.

[15]. W. Jiang, M. Zhuang, C. Xie, and J. Wu, "Sensing attribute weights: A novel basic belief assignment method"," *Sensors*, vol. 17, no. 4, p. 721, 2017.

[16]. C. Holdgraf- *Case study 7: Feature extraction and data wrangling for predictive models of the brain in python", The Practice of Reproducible Research*. University of California Press, pp. 139-148, 2017.

[17]. C. Iwendi, A.K. Bashir, A. Peshkar *et al.*, "COVID-19 patient health prediction using boosted random forest algorithm"," *Front. Public Health*, vol. 8, p. 357, 2020.

[18]. V. V. Ramalingam, A. Dandapath, and M. K. Raja, "heart disease prediction using machine learning techniques: a survey"," *Int. J. Eng. Technol*, vol. 7, no. 2, pp. 684–687, 2018.

[19]. B. Ambale, B. Venkatesh, X. Yang *et al.*, "Cardiovascular event prediction by machine learning: the multi-ethnic study of atheroscle- rosis"," *Circ. Res*, vol. 121, no. 9, pp. 1092–1101, 2017.

[20]. F. Mehmood, H. U. Rashidkayani, and F. Hussain, "Chronic diseases modelling-python environment," *J. Biol*, vol. 10, no. 1, pp. 31–38, 2020.

[21]. M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature selection using random forest classifier for predicting prostate cancer"," *IOP Conf. Ser.: Mater. Sci. Eng*, vol. 546, no. 5, pp. 1–9, 2019.

[22]. S. W. A. Rizvi, V. K. Singh, and R. A. Khan, "Revisiting Software Reliability Engineering with Fuzzy Techniques," in *IndiaCom-2016) Proc. of the 3rd IEEE Int. Conf. on Computing for Sustainable Global Development*, New Delhi, India, pp. 16–18, 2016.

[23]. H. B. Yadav and D. K. Yadav, "Early software reliability analysis using reliability relevant software metrics," *Int. J. Syst. Assur. Eng. Manag.*, vol. 8, no. S4, pp. 2097–2108, 2017.

[24]. S. W. A. Rizvi and R. A. Khan, "Maintainability Estimation Model for object-oriented software in design phase (MEMOOD)," Journal of Computing, vol. 2 no. 4, pp.26-32, 2010.

[25]. S. W. A. Rizvi and R. A. Khan, "A Critical Review on Software Maintainability Models," in *Proceedings of the Conference on*, 2009.

[26]. S. K. Khalsa, "A Fuzzified Approach for the Prediction of Fault Proneness and Defect Density," *Proceeding of World Congress on Eng*, vol. 1, pp. 218–223, 2009.

[27]. O. P. Yadav, N. Singh, R. B. Chinnam, and P. S. Goel, "A Fuzzy Logic based approach to Reliability Improvement during Product Development," *Reliability Engineering and System Safety*, vol. 80, pp. 63–74, 2003.

[28]. D. Yuan and C. Zhang, "Evaluation strategy for software reliability based on ANFIS," in *2011 International Conference on Electronics, Communications and Control (ICECC)*, pp. 3738-3741,2011.

[29]. D.K. Yadav, S.K. Charurvedi and R.B. Mishra, "Early Software Defects Prediction using Fuzzy Logic". International Journal of Performability Engineering, vol. 8 no. 4, pp. 399-408, 2012.

[30]. S. Aljahdali, "Development of Software Reliability Growth Models for Industrial Applications Using Fuzzy Logic," *Journal of Computer Science*, vol. 7, no. 10, pp. 1574–1580, 2011.

[31]. V. Cortellesa, H. Singh, and B. Cukic, "Early Reliability Assessment of UML Based Software Models," in *Proceedings of the 3rd International Workshop on Software and Performance*, pp. 302–309, 2002.

[32]. C. Wholin and P. Runeson "Defect Content Estimations from Review Data," in *Proceedings of 20th International Conference on Software Engineering*, pp. 400–409, 1998.

[33]. H. B. Yadav and D. K. Yadav, "A Fuzzy Logic based Approach for Phase-wise Software Defects Prediction using Software Metrics," *Information and Software Technology*, vol. 63, pp. 44–57, 2015.

[34]. S. W. A. Rizvi, V. K. Singh, and R. A. Khan, "The state of the art in software reliability prediction: Software metrics and fuzzy logic perspective," in *Advances in Intelligent Systems and Computing*, New Delhi: Springer India, 2016, pp. 629–637.

[35]. S. Mohanta, G. Vinod, and R. Mall, "A technique for early prediction of software reliability based on design metrics," *Int. J. Syst. Assur. Eng. Manag.*, vol. 2, no. 4, pp. 261–281, 2011.

[36]. P. He, B. Li, X. Liu, J. Chen, and Y. Ma, "An empirical study on software defect prediction with a simplified metric set," *Inf. Softw. Technol.*, vol. 59, pp. 170–190, 2015.

[37]. M. Li and C. S. Smidts, "A ranking of software engineering measures based on expert opinion," *IEEE Trans. Softw. Eng.*, vol. 29, no. 9, pp. 811–824, 2003.

[38]. N. Martin, N. Fenton, and L. Nielson, "Building large-scale Bayesian networks," The Knowledge Engineering review, vol. 15, no. 3, pp. 257–284, 2000.