MILESTONE RESEARCH FOUNDATION
Int. Jrl. of Computational Learning & Intelligence
Vol. 02, Issue No. 01, 2023

ORIGINAL RESEARCH

Big Data Analytic approach to Predict Risk Assessment for Cardiovascular Diseases Using Framingham Risk Score

N Ch S N Iyengar[1] . T Vivekanandan[2] . Syed Thouheed Ahmed[3]

[1]Department of Information Technology
Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (T.S), India
[2]Depatment of Computer Science and Engineering,
Srinivasa Institute of Technology and Management Studies, Chittoor, AP, India
[3]School of Computing and Information Technology, REVA University, Bengaluru, India

Abstract – Big data analytics in healthcare is emerging as a promising field for providing imminent predictive analysis from very large data sets and improving outcomes while reducing costs. Its potential benefits are enormous. However there remain challenges to overcome. Cardiovascular diseases (CVD) are the major cause and threat of mortality globally, as well as in India. They are caused by disorders of the heart and blood vessels, and include heart attacks, stroke, raised blood pressure (hypertension), congenital heart disease and heart failure. Most of the CVD can be prevented through proper risk assessment and necessary measures. Prevention of CVD involves improving risk factors through healthy eating habits, physical activity, avoiding tobacco smoke and alcohol intake. Treating risk factors, such as high blood pressure, blood lipids and diabetes is also beneficial. In this paper, we used Framingham risk score for risk assessment and large scale data analytics using hadoop map-reduce programming to estimate the patient risk for the data taken from Sai Balaji Hospital, chittoor.

Index Terms – Big Data Analytics, Healthcare Analysis, Cardiovascular Diseases, Risk Assessment, Framingham Risk Score, cloud computing.

I. INTRODUCTION

Huge volume and complex form of data has been generated rapidly across the globe. The big challenge here is traditional data-processing software is inadequate to deal with them [15,16]. In particular the healthcare industry historically has generated huge volume of data, driven by record keeping, compliance & regulatory requirements, and patient care. The most of data are maintained through manually records [17]. The recent trend is in the direction of digitization the healthcare records, as result it is driven by mandatory requirements as well as to get better quality of potential healthcare delivery. This takes the major advantage of reducing the healthcare cost, advanced clinical decision support, handling massive volume of data, hold the promise of supporting complex medical and healthcare information, etc [18]. In recent years, healthcare data has reached nearly 150 exabytes. With this rate of growth, the healthcare data is expected to reach zettabyte (10^{21} gigabytes) soon and, not long after, the yottabyte (10^{24} gigabytes) of healthcare data [14].

Big data in healthcare is irresistible due large volume, diversity of data types and the rate of speed in which it is managed. The data from source related to patient make up big data in the healthcare sector. It includes clinical data like medical practitioner prescriptions, laboratory reports, medical images, pharmacy information and other administrative data maintained in Electronic Health Records (EHRs). The potential large healthcare analytics leads to better assessment of patient characteristics, recommendation of effective treatments, applying advanced analytics, taking proactive measures, etc.

Cloud computing is an on demand Information Technology (IT) based infrastructure allows to access highly reliable of computing resources. It offers different service models, through it allows to Infrastructure, Software and Platform as a services [5]. The adoption of cloud technology in healthcare organizations can offer various services effectively. Some of the services offered are cost effective services, time access to sensitive data, highly reliable storage, highly secured access, etc. Cloud computing support to manage broad network access and resource pooling from electronic medical records (EMRs) [6, 7].

The disorders of the heart and blood vessels are commonly referred as cardiovascular diseases (CVDs). It includes cerebrovascular disease, coronary heart disease, rheumatic heart disease and other conditions. About one third of deaths are due to CVD [11], in which heart attacks and strokes are about four out of five. Persons at risk are demonstrated due to smoking, diabetic, raised blood pressure, overweight and obesity, etc. These can all be easily estimated in primary healthcare services. Classifying those at utmost risk of CVDs and make sure they get proper treatment can avoid premature deaths. It is estimated that 90% of CVD is preventable. Prevention of CVD involves improving risk factors. Preventive Treatments for risk factors, such as high blood pressure, blood lipids and diabetes is also beneficial by systematic risk assessment. The Framingham risk score is a long-term, ongoing risk assessment, 10 years CVD risk of an individual can be estimated[1,2]. It is based on findings of the Framingham Heart Study.

In this paper, we collected the data from “Sai Balaji Hospital”, chittoor for risk assessment of cardiovascular disease. The patient risk assessment was estimated use Framingham risk score. Further, other risk factors have been analyzed with Map-reduce programming in cloud environment. The result has been discussed and visualized.

II. LITERATURE REVIEW

According to literature existing, about 90% of data has been generated within last two years. Huge volume of data generated through various sources like, web social media, healthcare sector, mobile networks, business sectors, etc. The generated data are in the form of structured, semi structured and unstructured. Analyzing and exploring of such records is highly inefficient through traditional processing system. Cardiovascular diseases (CVDs) are the major cause of death globally. Due enlarged globalization and urbanization, it is noticed that the risk factors for cardiovascular disease increased [4, 10]. About 17.7 million people die every year due CVDs and it estimated to be 31 % of all global deaths. Nearly 75 % CVD deaths occur in low income middle-income countries. Further it is expected that by 2030, 23.6 million people die due to heart attack and any other cardiovascular disease [13]. In India, particularly in Andhra Pradesh cases reported that between 21-25 % is at the age of 60 [12]. Further it is estimated most of the deaths are due to CVD and average age is about 49 years. It is also noticed that, highest number of cases are reported in chittoor, Guntur and krishna districts. The report states that 13,840 heart attack patients were hospitalized in the year 2016.

The risk assessment and prevention at the primary stage is simple to adopt and implement than secondary stage of prevention. The prevention of CVD at primary stage focuses on detection of high-risk factor that are the major cause of CVD. The key query to be resolved is how to define different risk levels and assess the CVD risk. In this work, we estimated the risk levels of CVD using Framingham Risk Score. For this work, we used various risk factors like age, diabetes, sex, blood pressure, LDL cholesterol, HDL cholesterol, Body Mass Index (BMI), smoking, etc. to estimate CVD risk for 10-years.

III. PRELIMINARIES

Framingham Heart Study

The Framingham Heart Study is a long-term heart study on residents of the town of Framingham, Massachusetts. The study was started in 1948 and commissioned with 5,209 adult subjects from Framingham, and at present third generation of participants is taking part in this long-term study [3]. Several literatures have been published related to Framingham Heart Study. It is widely accepted by many researchers as outstanding risk assessment tool for CVD [8, 9]. From the beginning of the Framingham Heart Study it was assumed that, cardiac health can be subjective by lifestyle and environmental factors, and also by genetic factors. This study is the source for the term risk factor. Earlier to the Framingham Heart Study, doctors had slight sense of prevention. Based on Framingham Risk Score, Framingham Heart Study is estimated for 10-year cardiovascular risk [19].

Framingham Risk Score

The Framingham Risk Score (FRS) is a long-term risk assessment algorithm. It one of the scoring systems used to identify the chances of individual's developing cardiovascular disease. In the recent days, several scoring systems are adopted for risk assessment. Cardiovascular risk Framingham scoring systems are estimated as the probability that an individual get developed cardiovascular disease in a specified time, generally 10 to 30 years [20,21]. FRS has been validated across the globe and there is only very minimum evidence for improvement beyond FRS.

IV. PROPOSED ALGORITHM

Risk assessment of Cardiovascular diseases using Framingham risk score in Big Data environment

In this work, we have adopted Risk Assessment of Cardiovascular Diseases using Framingham Risk Score in Big Data Environment and cloud. The CVD risk assessment and large data analytics is given in the Fig.(1). Large healthcare data are generated from various healthcare sources like patients medical records, laboratories, web portal, mobile healthcare applications, hospitals, pharmacies etc., in the form of EMR. Data maintained in the traditional storage imitates a great challenge for managing large volume of healthcare data. Hence traditional storage has been replaced by highly reliable Cloud storage. It supports to handling of huge dynamic data efficiently.

Large-scale and complex data processing is carried out using map reduce programming in Hadoop big data environment. It delivers effective services to take the complete benefit of big data. The potential benefits include effective decision support system, advanced prediction, risk assessment, patient care, etc. In this work, CVD risk assessment is carried out using Framingham risk score. Based on the risk assessment, effective clinical analysis and recommendation are carried out effective. Further, healthcare organization performs data visualization for further analysis and future direction.

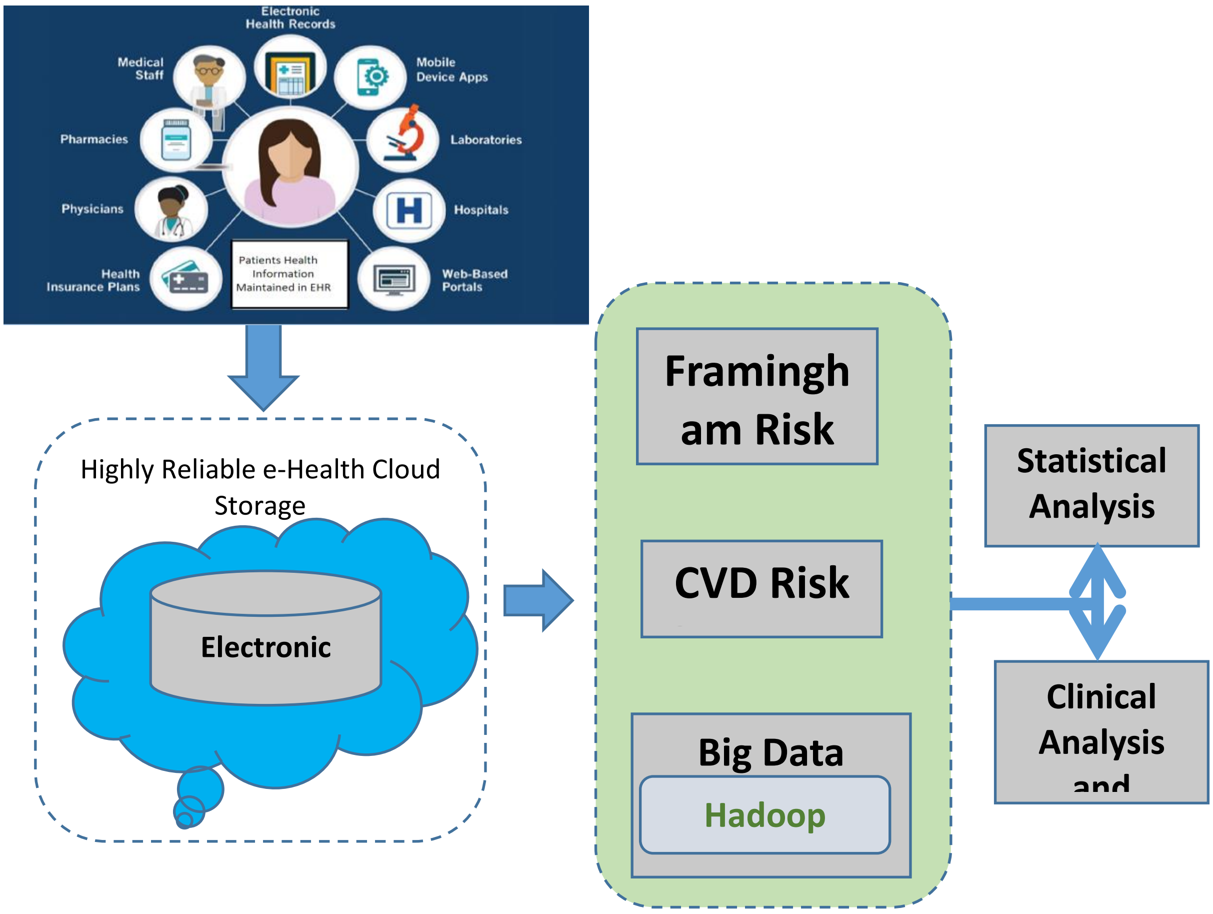

Fig.1.Road map cycle for designing and implementing collaborative learning.

Dataset

In our work, we implemented risk assessment model for CVDs using Framingham risk score algorithm and map-reduce program. For experimental analysis, dataset was collected from Sai Balaji Hospital, Chittoor. Further, Dataset has been pre-processed and handled for large scale data analytics. Dataset has been managed in cloud environment have been imported and processed to Hadoop-HDFS. Further, Large scale data analysis has been carried out using Map-Reduce programming and Framingham risk score algorithm for risk analysis and recommendation. For this study, Framingham risk score to predict long-term risk has been carried based on Framingham risk maintained by ST ALBANS and HEMEL HEMPSTEAD NHS TRUST, Cardiology Department.

V. RESULTS AND DISCUSSION

CVD risk analysis has carried out using Framingham risk and various risk levels has been identified and analyzed. High risk, Low risk, Medium risk, Very low risk levels has considered for this work.

Age based CVD Risk Analysis

Age based Risk analysis is mentioned in the Table.1 with various risk levels with varying age.

Table.1 – Age based CVD Risk Analysis

AGE	30-40	40-50	50-60	>60
HR	0	2	7	10
MR	0	5	7	1
LR	0	4	2	0
VLR	7	9	2	0

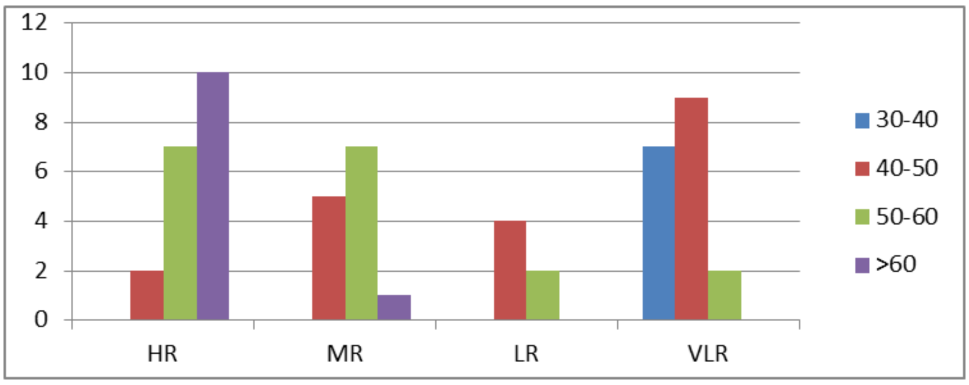

Fig.2. Age based CVD Risk Analysis

It is inferred from above Fig.2 that, there is higher chances of risk at older age compared to middle and younger age.

Diabetics based CVD Risk Analysis

Diabetic based Risk analysis is mentioned in the Table.2 with various risk levels with varying age.

Table.2 – Diabetics based CVD Risk Analysis

Diabetics	HR	MR	LR	VLR
Individuals without Diabetics	3	1	3	11
Individuals with Diabetics	6	1	0	2

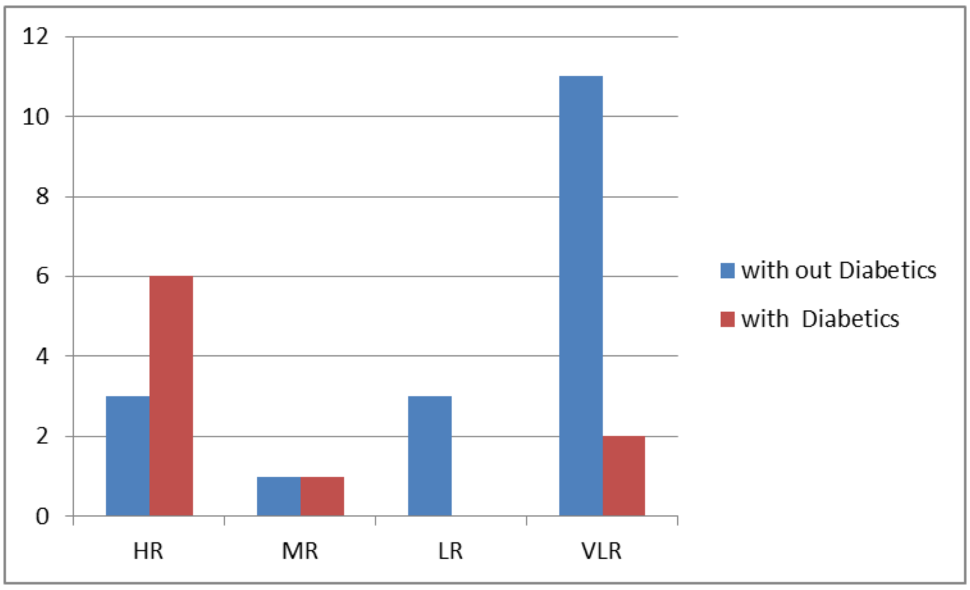

Fig.3. Diabetics based CVD Risk Analysis

It is observed from Fig.3 that, the patients with diabetics are having more chances of CVD risk compared with patients without having diabetics.

Smoking based CVD Risk Analysis

Various risk levels of smoker and Non-Smoker is mentioned in the Table.3

Table.3 – Smoking based CVD Risk Analysis

	HR	MR	LR	VLR
Non-Smoker	**3**	**5**	**4**	**9**
Somker	**13**	**6**	**2**	**3**

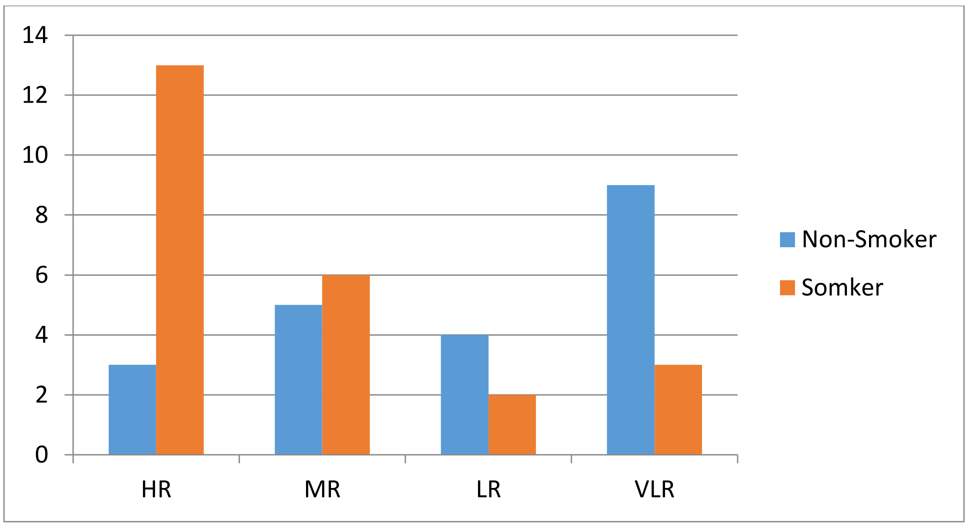

Fig.4. Smoking based CVD Risk Analysis

The smokers are at higher risk compared to non-smokers, which can be inferred from the Fig.4. Further, smoking habit can highly influence in getting CVD risk.

Cholesterol based CVD Risk Analysis

Various risk levels and different cholesterol range is given in the Table.4

Table.4 – Cholesterol based CVD Risk Analysis

Cholesterol	0 to1	1 to 1.5	1.5>
HR	**7**	**3**	**9**
MR	**4**	**6**	**4**
LR	**3**	**4**	**1**
VLR	**5**	**11**	**1**

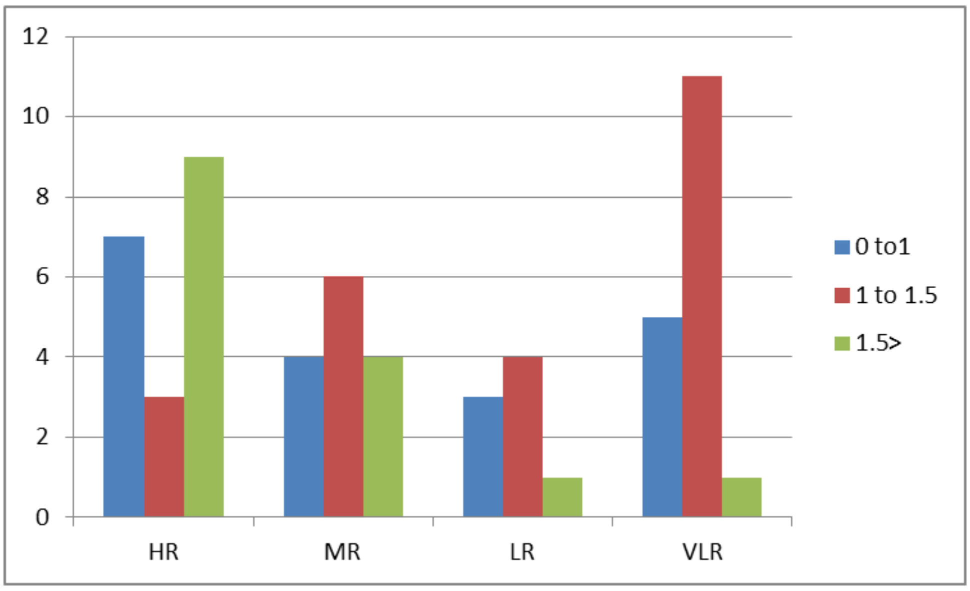

Fig.5. Cholesterol based CVD Risk Analysis

It is observed in Fig.5 that, higher cholesterol values have more chances of CVD risk compared with patients having lower range.

VI. CONCLUSION AND FUTURE WORK

The HealthCare sources generate large volume data and expected to generate at exponential rate in the years to come. The deaths across the globe due to CVD are a major concern as it is about one-third of the deaths globally and most of which may be prevented. Hence, it is a prime concern for healthcare sector to develop a risk assessment model to perform primary different analysis on Large Healthcare data maintained in Cloud environment. In this paper, the risk assessment model has been developed for CVD using Framingham risk score algorithm. For analysis, dataset collected from Sai Balaji Hospital, Chittoor has taken for study. Various analyses have been carried out to demonstrate CVD risk and its impact. The experimentation has been carried out using map reduce programming in Hadoop Big Data environment. In the present study we are greatly limited with data size. In future work, we planned to extend our analysis with the large volume as well different regions across the state.

REFERENCES

1. Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, *383*(9921), 999-1008.
2. Pencina, M. J., D'Agostino Sr, R. B., Larson, M. G., Massaro, J. M., & Vasan, R. S. (2009). Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. *Circulation*, *119*(24), 3078-3084.
3. Ahmed, S. T., & Syed Mohamed, E. (2021). Phonocardiography (PCG) Signal Optimization and Compression for Low Line Transmission in Telemedicine. In *Advances in Automation, Signal Processing, Instrumentation, and Control* (pp. 1127-1137). Springer, Singapore.
4. Kannel, W. B., & McGee, D. L. (1979). Diabetes and cardiovascular disease: the Framingham study. *Jama*, *241*(19), 2035-2038.
5. Lu, G., & Zeng, W. H. (2014). Cloud computing survey. In *Applied Mechanics and Materials* (Vol. 530, pp. 650-661). Trans Tech Publications Ltd.
6. Liu, Z., Weng, J., Li, J., Yang, J., Fu, C., & Jia, C. (2016). Cloud-based electronic health record system supporting fuzzy keyword search. *Soft Computing*, *20*(8), 3243-3255.
7. Rimal, B. P., Choi, E., & Lumb, I. (2009, August). A taxonomy and survey of cloud computing systems. In *2009 Fifth International Joint Conference on INC, IMS and IDC* (pp. 44-51). Ieee.
8. Mendis, S. (2010). The contribution of the Framingham Heart Study to the prevention of cardiovascular disease: a global perspective. *Progress in cardiovascular diseases*, *53*(1), 10-14.
9. Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., ... & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158-164.
10. 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. *Computers in biology and medicine*, *90*, 125-136.
11. Thanigaivasan, V., Narayanan, S. J., Iyengar, S. N., & Ch, N. (2018). Analysis of parallel SVM based classification technique on healthcare using big data management in cloud storage. *Recent Patents on Computer Science*, *11*(3), 169-178.

12. Prabhakaran, D., Jeemon, P., & Roy, A. (2016). Cardiovascular diseases in India: current epidemiology and future directions. *Circulation*, *133*(16), 1605-1620.
13. Al-Shammari, N. K., Alzamil, A. A., Albadarn, M., Ahmed, S. A., Syed, M. B., Alshammari, A. S., & Gabr, A. M. (2021). Cardiac Stroke Prediction Framework using Hybrid Optimization Algorithm under DNN. *Engineering, Technology & Applied Science Research*, *11*(4), 7436-7441.
14. Prasad, P. D., Vivekanandan, T., & Srinivasan, A. (2015). A Methodology for WebLog Data analysis using HadoopMapReduce and PIG. *i-manager's Journal on Cloud Computing*, *3*(1), 13.
15. Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big data analytics and big data science: a survey. *Journal of Management Analytics*, *3*(1), 1-42.
16. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.
17. Spruit, M., & Lytras, M. (2018). Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. *Telematics and Informatics*, *35*(4), 643-653.
18. Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological forecasting and social change*, *126*, 3-13.
19. Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino Sr, R. B., ... & Benjamin, E. J. (2009). Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *The Lancet*, *373*(9665), 739-745.
20. Thanassoulis, G., Peloso, G. M., Pencina, M. J., Hoffmann, U., Fox, C. S., Cupples, L. A., ... & O'Donnell, C. J. (2012). A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circulation: Cardiovascular Genetics*, *5*(1), 113-121.
21. LK, S. S., Ahmed, S. T., Anitha, K., & Pushpa, M. K. (2021, November). COVID-19 Outbreak Based Coronary Heart Diseases (CHD) Prediction Using SVM and Risk Factor Validation. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-5). IEEE.
22. Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, *383*(9921), 999-1008.