INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS
VOLUME XVI, ISSUE V, MAY 22, WWW.IJCEA.COM, ISSN 2321-3469

DESIGNING DISEASE PREDICTION MODEL USING MACHINE LEARNING APPROACH

Prof. N.R. Jain, Harshal Chaudhari, Lalit Jadhav and Vaishnavi Patole

Department of Information Technology, PDEA's COE, Manjari, Hadapsar. Maharashtra, Pune 412307

ABSTRACT:

Nowadays, human creatures confront different afflictions since of the natural circumstance and their dwelling conduct. So the expectation of affliction at an in progressed degree will got to be a crucial assignment. But the right forecast on the thought of signs will ended up as well difficult for specialists. The exact forecast of ailment is the greatest difficult errand. To triumph over this inconvenience, data mining performs a crucial work to are anticipating the sickness. Therapeutic mechanical know-how contains a enormous amount of data increment concurring to year. Due to the increased amount of data increment withinside the clinical and healthcare range the proper assessment of clinical data has been cashing in on early influenced individual care. With the help of ailment data, data mining uncovers covered up test truths in a expansive amount of clinical data. We proposed in vogue ailment expectation fundamentally based completely at the signs of the influenced individual. For the affliction expectation, we utilize K-Nearest Neighbor (KNN) and Convolutional neural organize (CNN) contraption examining set of rules for the proper forecast of ailment. For affliction forecast required affliction signs dataset. In this in vogue ailment forecast, the dwelling conduct of somebody and checkup facts do not disregard for the right expectation. The precision of elegant ailment forecast through way of implies of the utilize of CNN is 84.5% that's additional than the KNN set of rules. And the time and the memory prerequisite also are additional in KNN than CNN. After in vogue ailment forecast, this device is prepared t provide the peril related to a elegant affliction that's a diminish peril of in vogue ailment or higher.

Key Words: CNN, KNN and Machine learning, Disease Prediction.

1. INTRODUCTION

With a developing slant of inactive and need of physical exercises, maladies related to liver have gotten to be a common experience these days. In provincial ranges the concentrated is still reasonable, but in urban ranges, and particularly metropolitan regions the liver infection may be a

exceptionally common locating these days. Liver maladies cause millions of passings each year. Viral hepatitis alone causes 1.34 million passings each year. Issues with liver patients are not easily found in an early organize because it will be working regularly indeed when it is mostly harmed. An early determination of liver issues will increment patient's survival rate. Liver disappointments are at tall rate of chance among Indians. It is anticipated that by 2025 India may gotten to be the World Capital for Liver Illnesses. The broad event of liver contamination in India is contributed due to deskbound way of life, expanded liquor utilization and smoking. There are around 100 sorts of liver diseases.

3 Major diseases are studied in given research project work.
1) Liver disease prediction
2) Breast cancer Prediction
3) Diabetes prediction

- The liver is an monstrous, noteworthy organ within the human body. Weighing around 3 pounds. The liver contains two tremendous parcels, called the benefit and the cleared out projections. The gallbladder sits beneath the liver, adjacent parts of the pancreas and stomach related organs. The liver and these organs coordinate to prepare, ingest, and prepare food. The liver's principal work is to channel the destructive substances within the blood beginning from the stomach related system, some time recently passing it to anything is cleared out of the body. No chance is however known to create up for the nonappearance of liver capacity within the long pull, yet liver dialysis methods can be utilized incidentally. Manufactured livers are however to be made to progress long pull substitution without the liver. Beginning at 2017, liver transplantation is the most elective for wrap up liver dissatisfaction. Liver hurt is the one of the most excellent deadliest afflictions on the planet. The elemental driver of liver hurt are Greasy liver, Liver Fibrosis, Cirrhosis, hepatitis and illnesses. Illustrates the stages of liver hurt, within the central orchestrate strong liver will conclusion up oily liver since of gathering of cholesterol and triglycerides, taking after couple of months to a long-time oily liver will ends up liver fibrosis, afterward it prompts final stage of liver hurt known as cirrhosis. Within the starting times of the liver sickness, it is outstandingly difficult to distinguish in spite of the reality that liver tissue has been hurt nicely, it sources various therapeutic masters over and over disregard to analyze the affliction. This could turn to off-base pharmaceutical and treatment, so early area is basic and vital to save the quiet.
- Breast cancer (BC) is one of the foremost frequent threatening tumors within the world, bookkeeping for 10.4% of all cancer passings in ladies matured between 20 and 50. Agreeing to the World Wellbeing Organization figures, 2.3 million ladies will be analyzed with BC in 2020. BC has been analyzed in 7.8 million women within the past 5 a long time, making it the foremost visit danger around the world. BC causes more disability-adjusted life a long time (DALYs) in ladies around the world than any other sort of cancer. BC strikes ladies at any age after adolescence in each country on the planet, with rates rising as they gotten to be more seasoned. For all of these reasons, there's an continuous require for a dependable and exact framework that can be utilized to assist within the early location and conclusion of BC infections to decrease the number of passings. Within the field of restorative investigation, machine-learning (ML) calculations can be connected broadly, for illustration, anticipating COVID-19, foreseeing Alzheimer's movement, foreseeing constant maladies, foreseeing liver clutter, heart illness, cancer, and others. ML and profound

learning (DL) play a critical part in tackling wellbeing issues and distinguishing infections, such as cancer expectation. Numerous analysts have connected ML and DL methods to create models and frameworks to foresee BC. Healthcare segments have huge volume databases. Such databases may contain organized, semi-structured or unstructured information. Big data analytics is the method which examinations colossal information sets and uncovers covered up data, covered up designs to find information from the given data.

- Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million.Diabetes Mellitus (DM) is classified as-

Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A procedure called, Prescient Examination, joins a assortment of machine learning calculations, information mining methods and measurable strategies that employments current and past information to discover information and foresee future occasions. By applying prescient investigation on healthcare information, critical choices can be taken and forecasts can be made. Prescient analytics can be done utilizing machine learning and relapse strategy. Prescient analytics points at diagnosing the malady with best conceivable exactness, upgrading persistent care, optimizing assets at the side progressing clinical outcomes. Machine learning is considered to be one of the foremost critical manufactured insights highlights bolsters improvement of computer frameworks having the capacity to obtain information from past encounters with no require of programming for each case. Machine learning is considered to be a desperate require of today's circumstance in arrange to kill human endeavors by supporting robotization with least imperfections. Existing strategy for diabetes discovery is employments lab tests such as fasting blood glucose and verbal glucose resilience. Be that as it may, this strategy is time expending. This paper centers on building prescient demonstrate utilizing machine learning calculations and information mining procedures for diabetes prediction.

2.LITERATURE SURVEY

M. Chen Proposed [1] a brand new multimodal disease hazard prediction set of rules primarily based totally on Convolutional Neural Network (CNN) with the aid of using the use of prepared and unorganized records of hospital. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang Discovered disorder prediction device for diverse regions. They achieved disorder prediction on 3 extraordinary illnesses inclusive of diabetics, cerebral infraction and coronary heart disorder. The disorder prediction is achieved on prepared records. Prediction of coronary heart disorder, diabetes and highbrow infraction is achieved with the aid of using the use of diverse system mastering set of rules like naïve bayes, Decision tree and KNN set of rules. The final results of Decision tree set of rules plays higher than KNN set of rules and Naïve bayes. Also, they predict that both a affected person enjoy from the excessive hazard of cerebral infarction or minimal hazard of

cerebral infarction. They used CNN primarily based totally multimodal disorder hazard prediction on textual content records, for the hazard prediction of cerebral infraction. The accuracy contrast takes region among CNN primarily based totally unimodal disorder hazard predictions against CNN primarily based totally multimodal disorder hazard prediction set of rules. The accuracy of disorder prediction resulted as much as the 94.8% with greater speedy velocity than CNN primarily based totally unimodal disorder hazard prediction set of rules. Step of comparable as that of the CNN-UDRP set of rules the CNN primarily based totally multimodal disorder hazard prediction set of rules step simplest the testing steps carries of extra steps. Given paper paintings on each the form of dataset like prepared and unorganized records. Author labored on unorganized records. While preceding paintings simplest primarily based totally on prepared records, none of the writer labored on unorganized and semi- prepared records. But this device proposed paintings is relying on prepared in addition to unorganized records. B. Qian, X. Wang, N. Cao, H. Li, and Y.- G. Jiang [2] deliberate the Alzheimer disorder hazard prediction device with the help of EHR data of the affected person. Here they used energetic mastering context to address a authentic problem persisted with the aid of using the affected person. In this the hazard version changed into construct. For that energetic hazard prediction set of rules is used the hazard of Alzheimer disorder. IM. Chen, Y. Mama, Y. Li, D. Wu, Y. Zhang, and C. Youn [3] proposed wearable 2.zero device wherein configuration eager cleanable material that improves the QoE and QoS of the next-technology healthcare device. Chen dependent new IoT primarily based totally records series device. In that new sensor primarily based totally clever cleanable garments created. By the applied of this garments, professional stuck the affected person physiological condition. What's greater, with the help of the physiological records evaluation occur. In this reversal of cleanable clever material consisting of a couple of sensor, wires and cathode with the help of this component factor person can geared up to collect the physiological nation of affected person in addition to emotional fitness repute data used of cloud primarily based totally device. With the help of this material, it stuck the physiological nation of the affected person. Also, for the exam reason, this data is applied. Examined the troubles which might be confronting even as designing wearable 2.zero architecture. The troubles in present device encompass physiological records amassing, poor intellectual impacts, antiwireless for frame quarter networking and Sustainable massive physiological records accumulation and so on. The severa sports achieved on statistics like exam on records, tracking and prediction. Again author classify the useful additives of the clever apparel representing Wearable 2.zero into sensors Integration, electrical-cable-primarily based totally networking, virtual modules. In this, there are various packages pointed out like continual disorder tracking, aged humans care, emotion care etc. Y. Zhang, M. Qiu, C.- W. Tsai, M. M. Hassan, and A. Alamri [4] designed cloud-primarily based totally fitness –Cps device wherein offers with the massive degree of biomedical records. Y. Zhang tested big degree of data improvement withinside the medicinal field. The data is made in the much less degree of time and the ordinary for data is positioned away in diverse configuration so that is the issue that the problem recognized with the massive records. In this designed the Health-Cps device in that improvements lean one is cloud and 2d one is massive records technology. Cloud-like records evaluation, tracking and prediction of records. With the help of this device, an character receives greater records approximately a way to cope with and cope with the great degree of biomedical data withinside the cloud. The 3 layers remember records series layer, records control layer and dataoriented layer. The records amassing layer positioned away allotted garage and parallel computing. The records control layer used for allotted garage and parallel computing. By this framework diverse obligations are completed with the help of Health-cps gadget, the numerous Health-cps systems. Related to healthcare know through this gadget. L. Qiu, K. Gai, and M. Qiu in [5] proposed telehealth gadget and tested a way to address plenty of health facility information withinside the cloud. This paper creator proposed strengthen withinside the telehealth gadget, that is for the most element depending on the sharing information amongst all of the telehealth offerings over the cloud. Yet, the data sharing at the cloud confronting hundreds of problems like community capability and digital system switches. In this proposed the information sharing on

cloud technique for the higher sharing of data via the information sharing ideas. Here deliberate the appropriate method for telehealth sharing model. this model, creator consciousness on transmission probability, community abilties and timing constraints. For this writer concocted new huge information sharing set of rules. By this calculation, customers get the appropriate association of coping with biomedical information. Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh [6] proposed a great scientific selection-making gadget which predicts the disorder primarily based totally on historic information of patients. In this anticipated numerous sicknesses and inconspicuous instance of affected person condition. Designed a great scientific selection- making gadget applied for the genuine disorder prediction at the historic information. In that moreover determined numerous sicknesses idea and hid instance. For the notion purpose on this used 2D/three-D graph and pie Charts.And 2D/three-D graph and pie charts illustration purpose. S.Leoni Sharmila, C.Dharuman and P.Venkatesan [13] offers a comparable research of numerous system getting to know method such Fuzzy logic, Fuzzy Neural Network and selection tree. They remember information set to categorise and do examine approximately nearly. As indicated through examine Fuzzy Neutral Network offers 91 %Acuracy for category in liver disorder dataset than different system getting to know set of rules. Author applied Simplified Fuzzy ARTMAP in modified nature of utility domain names and is succesful to carry out category all round productively and giving enormously advanced performances. Author have reasoned that system getting to know algorithms for instance, Naive Bayes and Apriori [14] are very precious for disorder analysis at the given information set. Here little extent information applied for prediction like signs and symptoms or beyond getting to know were given from the bodily analysis. Confinement of this paper they could not remember big dataset, currently a day's medicinal information is growing so wishes to categorise that and category of that data is challenging. Shraddha Subhash Shirsath [15] proposed a CNN-MDRP set of rules for a disorder prediction from an big extent of health facility's prepared and unstructured data. Utilizing a system getting to know set of rules (Neavi- Bayes) Existing set of rules CNN-UDRP simply makes use of an prepared data but in CNN-MDRP middle round each prepared and unstructured data the accuracy of disorder prediction is greater and quick while contrasted with the CNN-UDRP. Here they consider remember huge information.

3.SYSTEM ARCHITECTURE

3.1 Block diagram

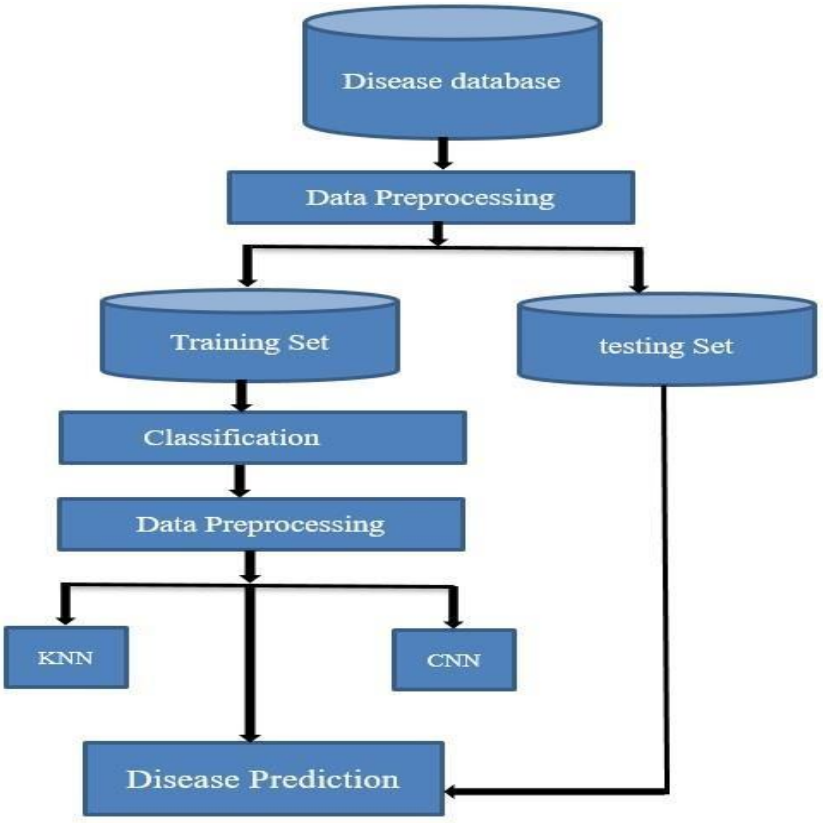


Fig. 1. Block diagram

3.2 At first, we take sickness dataset from UCI contraption considering web location and this is often inside side the shape of affliction posting with its indications. After that preprocessing is accomplished on that dataset for cleansing which is putting off commas, accentuations, and white places. And typically utilized as an instruction dataset. After that characteristics are extracted and chosen. At that point we classify that information the usage of type strategies alongside KNN and CNN. Based on contraption considering we'll anticipate the proper affliction. Algorithms and methods

1) K-Nearest neighbor (KNN)

1. Initially, we select a value for K in our KNN algorithm.
2. Now we go for a distance measure. Let's consider Eucleadean distance here. Find the euclidean distance of k neighbors.
3. Now we check all the neighbors to the new point we have given and see which is nearest to our point. We only check for k-nearest here.
4. Now we see to which class there is the highest number obtained. The max number is chosen and we assign our new point to that class.
5. In this way, we use the KNN algorithm.

2) Convolutional neural network (CNN)

1. The dataset is transformed into the vector form.
2. Then phrase embedding is done which undertakes zero values to fill the data. The output of phrase embedding is a convolutional layer.
3. This Convolutional layer is taken as entering to pooling layer and we carry out the max-pooling operation on the convolutional layer.
4. In Max pooling, the dataset is converted into constant length vector form. The pooling layer is attached with the complete related neural network.
5. The complete connection layer related to the classifier is the softmax classifier.

4. RESULT AND DISCUSSIONS

A. Experimental Setup:

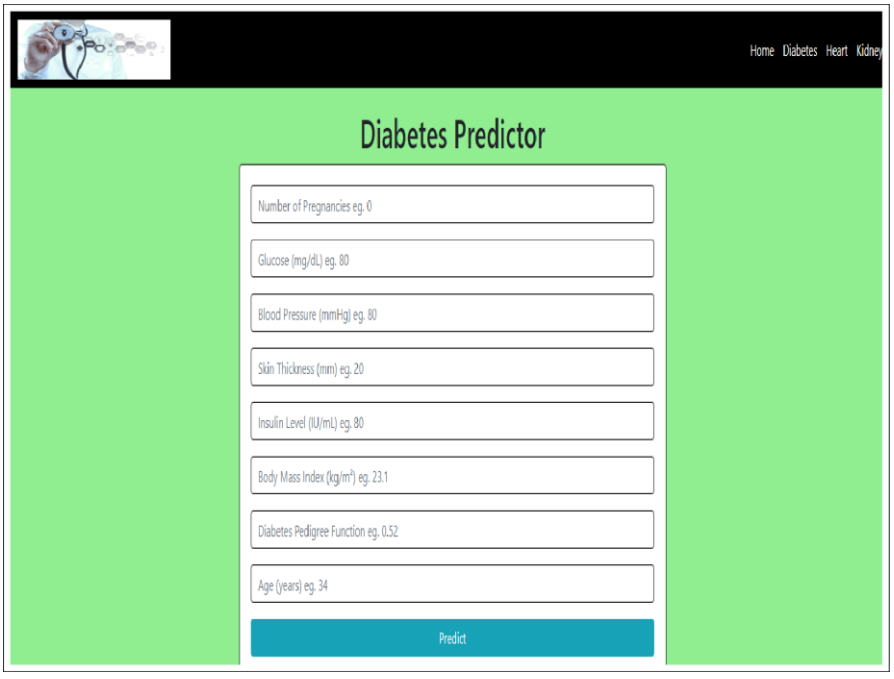
All the experimental instances are carried out in Python in congestion with Flask equipment and python interpreter as backend, algorithms and strategies, and the competing type technique at the side of diverse feature extraction technique, and run in surroundings with System having the configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (sixty- four bit) system with 8GB of RAM

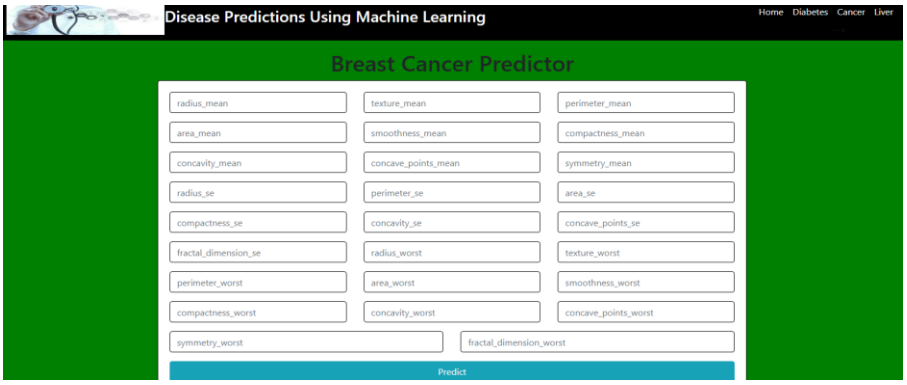
B. Dataset:

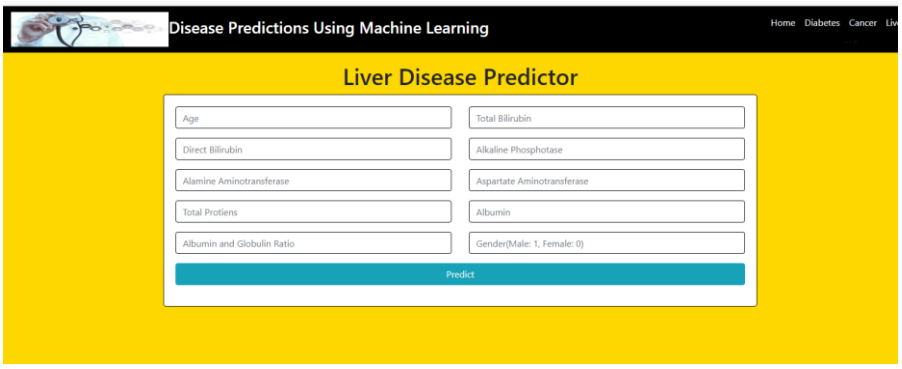
Patient disease dataset downloaded from UCI or Kaggle machine learning website.

C. Results:

This segment provides the overall performance of the KNN and CNN class algorithms in phrases of time required and reminiscence and different overall performance measures which include FP measure, precision, recall, and accuracy.







5. SOFTWARE

Python IDE: PyCharm / Jupiter

notebook Python

Interpreter/compiler: python 3.8/3.9

6. CONCLUSION

We proposed a malady expectation machine essentially based completely on side effects. For malady expectation basically based completely on indications, we utilized a contraption considering set of rules usually KNN and CNN. We carried out infection expectation through the utilize of the KNN set of rules and CNN set of rules. We assess the results among the KNN and CNN set of rules and the exactness of the CNN set of rules is 94% which is additional than the KNN set of rules. We were given exact infection forecast as yield, through giving the enter as sufferers record which help us to secure the degree of forecast. This machine may moreover lead to moo time admissions and minimal expense feasible for disease forecast. Within the future, we are able transfer additional afflictions and are anticipating the chance which influenced individual endures from the exact disorder.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang,"Disease prediction by machine learning over big data from healthcare communities", ," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.
[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Common., vol. 55, no. 1, pp. 54–61, Jan. 2017.
[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.
[5] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184–189.
[6] Disease and symptoms Dataset –www.github.com.
Heart disease Dataset-WWW.UCI Repository. com
[7] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," in IEEE big data analytics and computational intelligence, Oct 2017 pp.2325.
[8] Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.
[9] Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors", IEEE Conference on Information & Communication Technologies (ICT), vol., no., pp.1227-31,11- 12 April 2013.
[10] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
[11] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
[12] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
[13] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.
[14] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[15] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
[16] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
[17] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[18] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
[19] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[20] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.