International Journal of Research in Engineering, Science and Management
Volume 5, Issue 5, May 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

A Hybrid Approach on Smart Health Prediction using Data Mining

N. Sushma[1*], S. S. Greeshma[2], Sadanala Manasa[3], S. V. Bhaskar[4], Anidha Arulanandham[5]
[1,2,3,4,5]*Department of Computer Science, New Horizon College of Engineering, Bengaluru, India*

***Abstract*: The digital technology era demands the world to provide an excellent health system, in order to ensure the community to be alive and healthy. Objectives of this research paper is admin can login using his credentials, add new doctor details, add disease and its symptoms and manage data. Doctor can login with his credentials, and view appointment of patients. New user can sign up, they can login using user Id and password. Disease prediction is done when user enter symptoms. User can upload the reports. Chat instantly with doctor, they can book appointments and can give feedback about doctors. This study can be used for the data mining techniques such as medical field, research field, and educational field and various aspects. Due to the availability of computers and other regulations, huge amount of data is becoming available in medical and healthcare areas. As per the modern technology huge improvement has been made in computer field and therefore there is no need to deal with such a large amount of data at a same time. A major objective of this study is to evaluate data mining technologies in medical and healthcare applications to develop an accurate disease prediction. It is an amazing innovation which is of exorbitant interest in the current PC world. It is a sub area of PC sciences which utilizes previously existing information in different data sets to change it into new arrangement of results. It makes use of Deep learning, machine learning and database management techniques to extract new patterns from large data sets and the knowledge associated with these patterns. By using this technique data can be extracted automatically or semi automatically. The various parameters included in data mining are classifying, clustering and predictive analysis.**

***Keywords*: Classifier, Clustering, Data mining, Decision Tree, KNN, Predictive analysis, Regression, Smart prediction.**

1. Introduction

Data mining is the method involved with dissecting an enormous volume of data to observe patterns and examples. In clinical and medical care regions, because of guidelines and because of the accessibility of PCs, a lot of information is opening up. Such a lot of information can't be handled by people in a short measure of time to make determination, and therapy plans. A significant target is to assess information mining strategies in clinical and medical services applications to foster precise choices [13].

It involves analysing certain amount of data collected through various sites to locate certain patterns of occurrence to predict future tendencies of diseases, using several processes of effective data collection, warehousing, computer processing the disease predicted, displaying it to the user and redirect to the doctor accordingly. To predict disease this application is used for the healthcare development, patient care assistance system and health related information. It helps in analysing the survivability of any disease. With this functionality therefore, it serves a great purpose when it comes to predicting people’s health diseases especially when we have huge amount of data available. This study will provide a great advantage in the healthcare industry as it may be used to manage patients on their current health issues [1].

There are currently a lot of health care institutions that has been developed such as hospitals and medical centres which are crucial to maintain and improve the health of the people around us. It is a basic necessity of giving proper health care especially for every one of us. For every illness and diseases that people are facing today and sometime in the future, it is because of these medical institutions and all the doctors who worked at these places that have made our lives physically and mentally better and also healthy. Although hospitals now are well-maintained with their staffs working, there are still some issues that exist which cause the staffs to make poor clinical decision that affects a patient’s health like lack of qualified doctors, unorganized health information and poor communications between doctors and patients [2].

2. Literature Review

Health Care Industry is rich in information and hence data mining has become a necessity in our daily lives. Its main purpose is to improve the current health of the people that we have shared and created. A health care institution such as hospitals or medical centres would essentially consist of many numbers of doctors who are well qualified and specialized on treating patients of their current illness that they have and to restore them back to healthy life [14].

The paper “An approach of Interactive solution for smart health prediction using data mining” [7] aims in developing a computerized system to check and maintain your health by knowing the symptoms. This system proposed study of huge datasets from various angles and obtaining gist of useful information. These methods are useful in detecting diseases and providing proper remedy for the same [4]. Aims to calculate various methods of data mining in applications to develop

[*]Corresponding author: sushmareddy1357@gmail.com

decisions and also to provide a detailed discussion about medical [6].

In this modern era, new technologies have been created and developed to improve people's daily life especially for health care. Doctors and nurses are now guided by smart health prediction system for storing medical information that may be used for research and diagnosis [17]. Few years ago, doctors were expected to use their own experience to handle every medical situation that each patient were facing every single day. Although their current methodology may have saved people's lives back then, they are still errors and wrong doings that have put human life on risk [3]. It is without a doubt a heavy burden for everyone especially the medical staffs to understand that a number of decisions could heavily affect other people's lives and health, it is also why such system itself proves to be vital on guiding medical staff to make a proper clinical decision to cure and restore the human health [10], [16].

A smart health disease prediction system is a healthcare system that helps patients to know the disease through the symptoms they are facing, if the symptom is severe, they are directed to consult the associated doctor. The patient can upload their medical reports, the system will be able to extract the numbers in the report. Thereafter, doctor can prescribe medications accordingly, if any test is required doctor can suggest [4].

This system will provide the guidance and information needed for doctors to diagnose patient on their illness and it will eliminate the difficulties that the doctors are facing, particularly in their clinical decision-making process[11], [15], [18].The system would require to gather a whole lot of medical data that are important to be used on predicting a patient's health status, these patterns of information will be analysed by using data mining techniques in order to find correlations and discover new pieces of information from unstructured data [5], [6]. By using data mining tools, it will not only be able to produce reliable results with less time consumption and complexity but also with smart and efficient decision-making and useful information [12].

3. Proposed System

It may have happened so often that somebody need specialists promptly, however they are not accessible because of some explanation. The proposed framework permits clients to get moment direction on their medical problems through a canny medical services framework on the web.

Smart Health Disease Prediction is a specialised framework used for improving the task of checking the user at the initial level and displays the possible disease. It starts with getting some data about the user like login credentials, or if it's first-time user, he has to register. The healthcare prediction system allows the user to share their experiencing symptoms and issues [29]. It then mines user symptoms to check and validate for various illness and based on input it predicts the disorder it feels user's symptoms are associated with. The user gets accessibility to the next convenient framework like chat with doctor, book an appointment and suggest nearby doctors with details [30].

4. Methods

This section explains the data mining processes and algorithms with its application in health prediction. It also analyses the prospects related to the application of data mining techniques in health prediction.

1) Classification

Classification represents a data mining technique that requires to collect various of information and data for their attributes in order to be analysed. Once the attributes have been identified, the data can be further categorized and managed [23].

2) The Random Forest Classifier

Random forest classifier, as the name suggests, contains a huge number of individual decision trees that will operate as an ensemble. Each individual tree in the random forest splits out a prediction class and the class with the majority of votes becomes our model's prediction and test class. Random forest fits a large number of decision tree classifiers on various sub-samples of the dataset and uses averaging technique to improve the predictive and effective accuracy and also control over-fitting modelling [7], [18], [21].

3) Clustering

Clustering is a data mining technique that requires identifying data that relates to another according to its differences and similarities. It relies on visual approach that shows distribution of data in relation for people to understand [19].

4) Regression

Regression techniques involves identifying and analysing the relationship between variables in a dataset. It is a technique that is used in aspects of data modelling. The relationship between variables may vary depending on its instances [20].

5) Sequential Pattern

Sequential pattern is a technique that focus on discovering similar patterns in a data transaction during certain periods. This technique is useful to uncover deviation in the data that is happening at regular intervals over time.

6) Prediction

Prediction simply involves analysing events that are in the past to predict future events. So historical data that has been kept is used for examination to gain some insight that might be useful to predict what will happen in the future [35]-[37].

The system uses Java platform which access data from the database and SQL query language to build and access the models. It is proven that Naïve Bayes could identify all the significant medical predictors, though their research stated that it can be further improved with more data sets and attributes were provided for testing. [8]

7) Decision Tree Algorithm

Decision Tree figuring has a spot with a gathering of directed learning calculations. The general point of view of using Decision Tree is to make a well-prepared set to demonstrate which can be used for estimation of target factors by taking in decision standards got from before data (preparing data) [26], [28].

In software engineering, Decision tree learning component uses a decision tree to go from discernments around a thing to choices about the thing's goal regard. It is one of the farsighted showing approaches used in experiences, data mining and AI [24].

The overall technique of the framework is as following advances:

1. The dataset is taken from the UCI AI archive and Kaggle and different sites and gathered from genuine information.
2. Pre-handling is trying not to miss esteem either substitution or eliminate missing worth from the dataset.
3. The information decreases process rehashes until getting superior execution exactness.
4. Model development is finished by utilizing KNN arrangement calculations.
5. Models are assessed by utilizing execution assessment methods.
6. In the wake of assessing model select the best presentation exactness.
7. Examinations and anticipate ongoing kidney infection informational collection by utilizing the chose model.

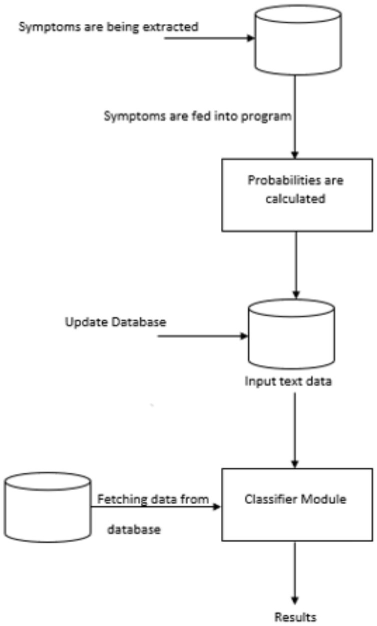

Symptoms are being extracted
Symptoms are fed into program
Probabilities are calculated
Update Database
Input text data
Fetching data from database
Classifier Module
Results

Fig. 1. Work flowchart

K-Nearest Neighbors: KNN portrayal is a champion among the most fundamental and clear game plan procedures and should be one of the primary choices for a request concentrate on when there is basically no prior finding out with regards to the spread of the data [27]. KNN request was created from the need to perform segregated assessment when trustworthy parametric appraisals of probability densities are dark or difficult to choose [25]. K-Nearest Neighbor is an otherwise called lethargic learning classifier. Characterization ordinarily includes parceling tests into preparing and testing classes [34].

Technologies utilized:

Eclipse IDE: For UI, Eclipse IDE (Integrated Development Environment) will be utilized for planning the Graphical User Interface (GUI).

Java: Java will be utilized for associating different parts of the UI to the database framework.

Navicat MySQL: MYSQL is utilized as a database at the web server. In this framework, server utilized is the tomcat serve Doctor, Patient and disease database is made with the assistance of Navicat MySQL. It gives a natural and ground-breaking graphical interface for database the board, improvement, and upkeep [38]-[40].

5. Comparison Analysis of Algorithms

Focused on a machine learning algorithm, we proposed a general method of disease prediction. We used Naïve Bayes algorithms to identify patient data because medical data are increasing in a vast rate, requiring the processing of existing data in order to predict exact disease based on symptoms. By having the input as a patient record and symptoms, we were able to get accurate and relatable general disease as a prediction that helped us knowing the degree of disease risk prediction [31]-[33]. Due to this method, disease prediction and risk prediction could be achieved over a short period of time and at a low cost [22]. In terms of accuracy and time, the results of Naïve Bayes and other algorithms are compared, and the accuracy of the Naïve Bayes algorithm is higher than the other algorithms.

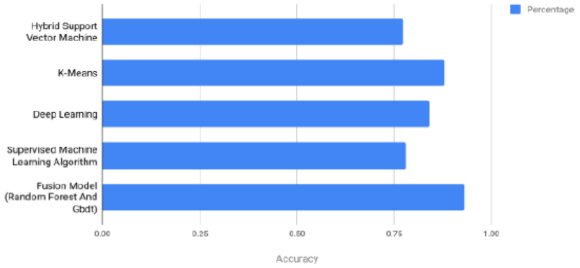


Fig. 2. Comparison of various techniques

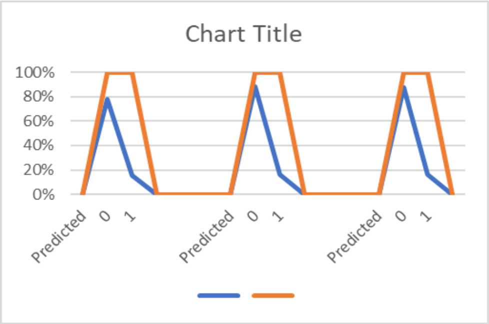
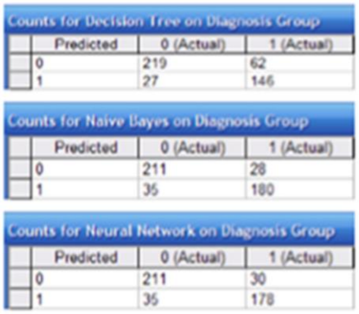


Fig. 3. Performance measure of proposed work



Counts for Decision Tree on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	219	62
1	27	146

Counts for Naive Bayes on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	211	28
1	35	180

Counts for Neural Network on Diagnosis Group

Predicted	0 (Actual)	1 (Actual)
0	211	30
1	35	178

Fig. 4. Results of classification matrix for all the models

6. Conclusion

The data mining can play a vital role in disease prediction to design a smart health prediction system. In medical diagnosis, data mining has been widely used for predicting diseases through diagnosis. However, no single data mining algorithm is best suited to resolve the prediction issues for healthcare data sets. In conclusion, the combination of several data mining or hybrid version of the data mining algorithm may be a better approach in designing health prediction system. The future research may be directed towards designing a better data mining-based model that can address healthcare with real-time healthcare datasets. This study does not encompass the complete analysis of all existing data mining algorithms and real-time healthcare dataset. Besides, the proposed health prediction system is not built through the comparison of all the data mining algorithms available in the literature. However, future research may be directed towards the selection of the best suitable data mining algorithm through the analysis of all existing algorithms.

7. Future Scope

Hidden knowledge will be extracted from the historical data in the proposed system, by preparing datasets by applying a hybrid approach. These datasets will be compared with the incoming queries and the final report will be generated using Association Rule Mining. Since this proposed methodology will work on real historical data, it will provide accurate and efficient results, which will help patients, get diagnosis instantly. More work can be done in the future by using more data set related to all diseases and by using different data reduction methods to improve the classification and clustering methods. For better accuracy and prediction of all diseases the datasets that will be used must be quality oriented and free from inconsistencies and missing values.

This web application can be further enhanced in an Android app. This will be available to users on mobile basis and its use can be further increased. Also, feature like getting the doctor online on chat so that patients can directly talk to the concerned doctors.

References

[1] R. Siddarth, I. Parvathi, “Survey on data mining techniques for diagnosis of diseases in medical domain”, International journal of computer science and information technologies, 2014.

[2] M. Tarawneh and O. Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," Acta Scientific Nutritional Health, vol. 3, pp. 147-151, 2019.

[3] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun et al., "Scalable and accurate deep learning with electronic health records", NPJ Digital Medicine, vol. 1, no. 1, pp. 18, 2018.

[4] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” 2017.

[5] Shubham Salunke, Shubham Rajiwade, Deepak Yadav, S. K. Sabnis, "Smart health prediction system using machine learning,” International Journal of Research and Analytical Reviews, vol. 7, no. 1, pp. 483-488, March 2020.

[6] Gupta A., Kumar L., Jain R., Nagrath P. (2020), Heart Disease Prediction Using Classification (Naive Bayes).

[7] Singh P., Pawłowski W., Tanwar S., Kumar N., Rodrigues J., Obaidat M. (eds), Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019). Lecture Notes in Networks and Systems, vol. 121. Springer, Singapore.

[8] U. Shruthi, V. Nagaveni and B. Raghavendra, "A review on machine learning classification techniques for plant disease detection", 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 281-284, 2019.

[9] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215.

[10] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, “Disease Prediction by Machine Learning Over Big Data from Healthcare Communities”, in IEEE Access, Special Sect ion on Healthcare Big Data, vol. 5, pp. 8869 – 8879, April 2018,

[11] B Geluvaraj, P. M Satwik, T. A. Ashok Kumar, Artificial intelligence (AI), and in particular machine learning (ML), deep learning (DL),739-747, Springer, Singapore.

[12] Anidha, M & Premalatha, K, “Integrated Cox model for survival analysis and biomarker discovery with a feature ranking technique based on z-score transformation in non-small cell lung cancer patients,” Biomedical Research, vol. 28, no. 5, pp. 1975-1983, 2017.

[13] Anidha, M & Premalatha, K, ‘An Application of Fuzzy Normalization in miRNA data for novel feature selection in cancer classification’, Biomedical Research, vol. 28, no .9, pp. 4187-4195, 2017.

[14] Anidha, M & Premalatha, K, “An mRMR with Mean Score Feature Selection for Ovarian Cancer Classification using Joint Analysis,” International Journal of Pharma and Bio Sciences, vol. 8, no. 2, pp. 495-504, 2017.

[15] Anidha, M & Premalatha, K, “An Entropy Based Mean Score Feature Selection Method for Identification of Biomarkers Using Mirna Expression Profiles for Cancer Classification,” Asian Journal of Information Technology, vol. 16, no. 2, pp. 206-211, 2016.

[16] Anidha, M & Premalatha, K, “A Hybrid Gene Selection Technique Using Improved Mutual Information and Fisher Score for Cancer Classification Using Microarrays,” World Academy of Science, Engineering and Technology, International Journal of Computer, Information, Systems and Control Engineering, vol. 10, no. 3, pp. 554-557, 2016.

[17] Uma. N, S. Prashanth CSR, “A detailed analysis of the various frequent itemset mining algorithms,” Journal of Advanced Research in Dynamical and Control Systems, vol. 12, Special Issue 2, pp. 448-454, 2020.

[18] Ilango V., Subramanian R., Vasudevan V, “Statistical data mining approach with asymmetric conditionally volatility model in financial time series data,” International Journal of Soft Computing, 2013.

[19] Kumar N., Sneha Y. S., Mungara J., Raghavendra Prasad S. G, “A Survey on Data Mining Methods Available for Recommendation System,” 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2017, 2018.

[20] Mohan Kumar S., Majumder D., Shajin Naragunam A., Ashoka D.V, A Symmetrically Diminished Interconnected Database Segmentation Framework Using Data Mining, Journal of Physics: Conference Series, Conference Paper, 1964, 2021.

[21] R Senthil Kumar, C. Ramesh, “Extreme Precipitation Events in Chennai Metro City Using Data Mining”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 11, September 2019.

[22] Khandelwal P., Gaspar F.B., Crespo M.T.B., Upendra R.S, Lactic acid bacteria: General characteristics, food preservation and health benefits, Fermented Foods: Part I: Biochemistry and Biotechnology, Book Chapter, 2016.

[23] Nithya B., Ilango V, Predictive analytics in health care using machine learning tools and techniques, Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017, January 2017.

[24] Adhikary P., Bandyopadhyay S., Kundu S, “Application of Artificial Intelligence in Energy Efficient H.V.A.C. System Design: A Case Study,” ARPN Journal of Engineering and Applied Sciences, 2017.

[25] Arulanandu C.K., Murthy S.D., Nagraj G, “Cloud based RDF security: A secured data model for cloud computing,” International Journal of Intelligent Engineering and Systems, 2018.

[26] Chalissery B.J., Asha V, “An intelligent organ distribution using Internet of Things-driven systems,” Proceedings of the 2nd International Conference on Communication and Electronics Systems, ICCES 2017, January 2018.

[27] Deshmukh V.M., Shukla S, Content-Restricted Boltzmann Machines for Diet Recommendation, Lecture Notes in Networks and Systems, Conference Paper, 290, 2021.

[28] Divya R., Chinnaiyan R, Reliable Smart Earplug Sensors for Monitoring Human Organs based on 5G Technology, Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018, 2018.

[29] Gautam K.S., Kaliappan V.K., Akila M, Strategies for Boosted Learning Using VGG 3 and Deep Neural Network as Baseline Models, Lecture Notes on Data Engineering and Communications Technologies, Book Chapter, 57, 2021.

[30] Godwin J.J., Krishna B.V.S., Rajeshwari R., Sushmitha P, Yamini M, IoT Based Intelligent Ambulance Monitoring and Traffic Control System, Intelligent Systems Reference Library, Book Chapter,193, 2021.

[31] Gopal M.K., Amirthavalli M, "Applying machine learning techniques to predict the maintainability of open-source software," International Journal of Engineering and Advanced Technology, 2019.

[32] Huchegowda C., Indumathi G., Huchegowda N, "Performance analysis of biorthogonal filter design using the lifting-based scheme for medical image transmission," International Journal of Computer Aided Engineering and Technology, 2021.

[33] Karthikayini T., Srinath N.K, Comparative Polarity Analysis on Amazon Product Reviews Using Existing Machine Learning Algorithms, 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2017, 2018.

[34] Muthuswamy J., "Extraction and classification of liver abnormality based on neutrosophic and SVM classifier," Advances in Intelligent Systems and Computing, 2019.

[35] Madhukar B.N., Jain S., Satyanarayana P.S, "Duality theorem for discrete Hartley transform," International Journal of Applied Engineering Research, 2015.

[36] Nirmala A.P., More S, "Role of artificial intelligence in fighting against covid-19," Proceedings of 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering, ICADEE 2020, 2020.

[37] Singhal R., Deepika N, "Detecting fraudulent words: Using PFCM clustering," 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 – Proceedings, 2017.

[38] Suma T., Murugesan R, "Artificial Immune Algorithm for Subtask Industrial Robot Scheduling in Cloud Manufacturing," Journal of Physics: Conference Series, 2018.