REVIEW ON AUTOMATIC DISEASE PREDICTION USING DATA MINING METHODS

Akhil Mathew Philip[1], *Dr. S. Hemalatha*[2]

ABSTRACT :

Heart disease (HD) has become a major cause of death killing 17.7 million people each year, 31 per cent of deaths in the world, according to the 2017 data of the World Health Organization. With expansion of the enormous dimensions of the dataset recently made available, HD inference can be conducted automatically using traditional empirical techniques to predict the potential of having HD on any person. This paper shows a programmed Heart Disease (HD) prediction strategy which depends on including strategies for evaluating and using data mining that provided side effects and clinical data in the patient dataset. The promising method for HD forecasting is knowledge mining, which allows the extraction of concealed skills from the data and investigates the link between qualities. The PC will learn viable HD indications to group HD into different classes. In any case, the data provided may include side effects that are unnecessary or interrelated. The use of such information could debase the execution of the grouping.

Keywords:

Data mining; Heart Disease Prediction; Feature Selection; Classificatioz

INTRODUCTION:

The clinical observation of healthcare diagnosis is usually a conclusion reached regularly by the information and practice of the specialist. PC Aided Decision Support System plays a major role in the field of restoration. Data mining offers the technique and creativity for turning these data hills into usable basic leadership data. By using data mining tools, some investment is needed to forecast the disease more accurately. Among the growing research on the anticipating structure of coronary disease, it has occurred to enormous classifications of the exploration results and gives per users with a design of the current methods of forecasting coronary disease in each category. Unfortunate propensities, such as tobacco use, unhealthy eating habit, physical inertia and alcohol consumption are the main reasons for many forms of HD. Including age, circulatory stress, cholesterol, hypertension, hyper pressure, a few clinical data and symptoms. HD database basically consists of the above-mentioned details and features that have been compressed and obtained from patients. Data mining devices may respond to exchange addressesthat are normally used for a long time to decide. Operating the HD software can be considered a true lifetime task and incorporating these apps helps patients understand the key risk factors associated with HD. With countless characteristics and by simply focusing on standard observable approaches, an attempt is made to identify what attributes are the most important risk factors for prediction of HD. This paper breaks down data analysis techniques that can be used to predict all kinds of diseases especially coronary disease, diabetes, malignant development in the breast, etc. In this Disease Prediction System, the client will include the Symptoms and the System will predict the possible occurrence of various sicknesses employing the data mining methods.

RESEARCH OBJECTIVES

- To study the significance of automatic disease prediction using data mining methods.

[1] *Research Scholar, Department of CS,CA & IT, Karpagam Academy of Higher Education, Coimbatore, India*

[2] *Assistant Professor, Department of CS, CA & IT, Karpagam Academy of Higher Education Coimbatore, India*

- To explore the existing prediction models for various diseases.

Literature Review: The Review of Literature includes various feature extraction and classification approaches used in automatic disease prediction systems.

- **Heart diseases**

Mohan, Thirumalai and Srivastava, 2019 proposed a new approach that involved the use of AI methods to boost precision in cardiovascular disease forecasting. Different blends of highlights and a few traditional classification schemes were introduced to the forecast model. Through the forecast model for coronary disease with the mix irregular timberland with a straight model (HRFLM), they produced an upgraded exhibition level with an accuracy of 88:7 per cent.

Ordonez, 2006 established a formula that used quest imperatives to decrease the number of values, scans for affiliation runs on a preparation set, and finally approved them on an autonomous test set. The scientific centrality of the identified values was tested with assistance, assurance and lifting. Principles of the association applied to a legitimate collection of information comprising rehabilitation reports for coronary disease patients. Jin et al., 2007 proposed a convincing and reliable cardiac breakdown prediction model. This paper's primary commitment was to predict cardiac breakdown using a neural network (i.e. to forecast the risk of heart disease based on the digital therapeutic information of the patient). Specifically, they used single-hot encoding and word vectors to display thedetermination occasions and predicted cardiovascular breakdown occasions using the basic standards of a long-term breakdown.

The genome-wide association reads for CVD results/qualities were illustrated by Pu, Zhao and Zhang, 2012. In general terms, clinical preliminaries on CVD expectations using genetic data were outlined. So far, the vast majority of single or specific genetic variants tested in subsequent clinical trials have not substantially improved CVD segregation. Papadaniil and Hadjileontiadis, 2013 presented a comprehensive heart sound division (HSS) technique that naturally identified the area of first (S1) and second (S2) heart sound and concentrated on from heart auscultative crude information. The heart phonocardiogram was broken down by using Ensemble Empirical Mode Decomposition (EEMD) in conjunction with kurtosis highlights to find the proximity of S1 and S2, and concentrate on it from the information reported, forming the proposed HSS plot, specifically HSS-EEMD / K.

- **Skin diseases**

Oladele, Olarinoye and Adebisi, 2018 offered a comparative evaluation of two knowledge mining characterization techniques, namely Decision Tree and Multi-layer Neural Network, which were applied for skin disease prediction. All research investigations were performed in the WEKA knowledge mining system context. Each individual classifier was prepared and evaluated using the N-overlap cross-approval protocol (N value was set to 10). The two classifiers were the independent family of Decision Tree and Neural Network. The prescient model acquired from the J48 and Multi-layer Perceptron (MLP) was estimated and evaluated with the use of essential parameters such as precision and kappa measurements as necessary. Li, Lin and Hwang, 2019 investigated the relationship between accommodating qualities and the occurrence of patients' weight wounds towards the end of their life. A retrospective study was conducted using information collected between January 2007 and October 2015 from 2062 patients towards the end of life. Given statistical information and weight damage scale assessment ratings, history of trauma, type of illness and duration of hospitalization were discovered as the significant autonomous factors for anticipating weight injury occurrence.

Verma, Pal and Kumar,2019 presented another strategy, which applied six different information mining arrangement systems and created a selection method usingBagging, AdaBoost and Gradient Classification procedures to predict skin disease groups. In addition, an item significance technique was used to pick the most striking 15 highlights that were expected to take on a significant job. Following the selection of the 15 highlights, a subset of the first data set was obtained to examine the consequences of six AI strategies, and an outfit approach was applied to the entire data set.

Incense et al., 20171 designed techniques that could be used to use current knowledge, obtained from liver digestion considerations, to inform skin digestion forecasts by understanding the differences between skin and liver in the enzymatic scenes. The device plot demonstrated how to use a variety of silico instruments to assess a vital test to predict risk after dermal introduction. The use of *in vitro* approaches to determine skin metabolism was also addressed in order to provide more exploratory information to explain and track expectations.

- **Cancer diseases**

Ahmad et al., 2018 refined the techniques and systems used in our mission, where the goal was to create calculations to assess whether or not a patient had or was likely to have malignant lung growth using data set images using knowledge digging and AI for order and review. Likewise, the determination technique was profoundly disturbing thinking about the detection of malignancy disease.

Behadili et al., 2019 investigated genuine information on breast cancer among Iraqi women; the data were physically collected from a few early-stage Breast Cancer hospitals in Iraq. Data mining techniques found hidden information, unexpected instances and new data set guidelines, which suggested a large number of features.

Strategies for discovering interesting examples and data mining techniques monitored the redundant and ultimately superfluous ascribes. Nonetheless, the dataset was managed through the Weka stage (The Waikato Climate for Information Analysis).

Johnston, Catton and Swallow, 2019 used camcAPP, cBioPortal, CRN and NIH NCI GDC information gateway to publicly-usable massive prostate malignant growth data sets. All transcripts were unequivocally linked to repeat encoding or mitosis regulation and cell cycle- related proteins. The top entertainer was BUB1, one of four main MIR145-3P microRNAtargets that were upregulated in hormone-delicate just like PCa safe for mutilation. SRD5A2 changed its progressively dynamic structure over testosterone and had been related to biochemical repeat in contrast.

Different arrangement systems were used in AbouElNadar and Saad to the group for different patients if the bosom disease was sporadic or non-repetitive. K-Nearest Neighbor (KNN), Decision Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Bagging, Voting and Random Forest (RF) procedures were the order methods used. The dataset was taken from the AI archive of the University of California Irvine (UCI), and exams were performed with knowledge mining apparatus of Waikato Environment for Knowledge Analysis (WEKA).

- **Diabetics diseases**

Different arrangement systems were used in (AbouElNadar and Saad) to the group for different patients if the bosom disease was sporadic or nonrepetitive. K-Nearest Neighbor (KNN), Decision

Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Bagging, Voting and Random Forest (RF) procedures were the order methods used. The dataset was taken from the AI archive of the University of California Irvine (UCI), and exams were performed with knowledge mining apparatus of Waikato Environment for Knowledge Analysis (WEKA). For comparing our findings and reports from other researchers, the Pima Indians Diabetes Dataset and the Waikato Framework for Information Analysis toolbox have been used. The end shows that the model obtained a forecast accuracy 3.04% higher than those of various analysts. Our model also ensured that the reliability of the dataset was sufficient.

Ding et al., 2019 proposed a diabetic-complexity perception paradigm based on an improved similarity inactive Dirichlet (seLDA) model. In particular, after information pre- processing, we first gaged the similarity between printed medicinal records and then per- structure seLDA-based diabetic difficulty theme mining depending on the requirements of closeness. Finally, by taking care of a multilabel grouping problem with help vectormachines (SVMs), we built a forecast model. The test results indicated that our approach outstripped the conventional LDA-based methodology by 22.49 per cent in similarity data.

Zhua, Idemudiaa and Feng, 2019 proposed PCA for dimensionality reduction, which characterized our dataset's reasonable starting centroids with k-implies calculation. Instead, K-implies was used to discover exceptions and package the knowledge into comparative meetings, with tactical relapse as a dataset classifier. Ahmad et al., 2018 refined the techniques and systems used in our mission, where the goal was to create calculations to assess whether or not a patient had or was likely to have malignant lung growth using data set images using knowledge digging and AI for order and review. Likewise, the determination technique was profoundly disturbing considering the detection of malignancy disease.

Behadili et al., 2019 investigated genuine information on breast cancer among Iraqi women; these data were physically collected from a few early-stage Breast Cancer hospitals in Iraq. Data mining techniques found hidden information, unexpected instances, and new data set guidelines, which suggested a large number of features.

Strategies for discovering interesting examples and data mining techniques monitored the redundant and ultimately superfluous attributes. Nonetheless, the dataset was managed through the Weka stage (The Waikato Climate for Information Analysis).

Johnston, Catton and Swallow, 2019 used camcAPP, cBioPortal, CRN and NIH NCI GDC information gateway to create publicly usable massive prostate malignant growth data sets. All transcripts were unequivocally linked to repeat encoding or mitosis regulation and cell cycle- related proteins. The top entertainer was BUB1, one of four main MIR145-3P microRNA targets that were upregulated in hormone-delicate just like PCa safe for mutilation. SRD5A2 changed its progressively dynamic structure over testosterone and was related to biochemical repeat in contrast.

Different arrangement systems were used in AbouElNadar and Saad to the group for different patients if the bosom disease was sporadic or non-repetitive. K-Nearest Neighbor (KNN), Decision Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Bagging and Voting and Random Forest (RF) procedures were the other methods used. The dataset was taken from the AI archive of the University of California Irvine (UCI), and exams were performed with knowledge mining apparatus of Waikato Environment for Knowledge Analysis (WEKA).

Comparison Analysis:

This section presents the current scenario of disease prediction models and find the research gaps

Reference No	Technique used	Benefits	Research gaps
1	The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM).	This improves the coronary disease prediction overview by providing a broader perspective.	In contrast to any constant applications, it is limited to hypothetical methodologies.
2	Association Rule Discovery with the Train and Test Approach	It provides high confidence, high lift.	? Large number of rules exist. ? Presence of validation rules on an independent rule.
3	An effective and robust Architecture using LSTM method	The records are managed sequentially.	Approach of expert knowledge is missing.
5	Heart sound segmentation (HSS) method that automatically detects the location of first (S1) and second (S2) heart sound	It is highly robust.	Murmur identification and extraction from the raw PCG is not carried out.
12	An approach for diabetic complication prediction depending on a similarity-enhanced latent Dirichlet allocation (seLDA) model	It surpasses the exhibition of the ordinary LDA approach and other seLDA-based methodologies	The time series information is missing.
13	A data-mining-based model for early diagnosis and prediction of diabetes using the Pima Indians Diabetes dataset	It achieves an enhanced k-means cluster result.	More improvement is needed in the model.

CONCLUSION:

Progressively predicting the course of events of persistent illness is a challenging job. It needs data about the specific disease context just as a pipeline that can redirect raw knowledge from an analysis process for general well-being into tailored infection forecasts. We have talked about preferences from a few distinctive empirical models, identifying places and periods in which our forecasts were correct. We also calculated the degree to which the predictions were constantly influenced by delayed reporting of events. Extensively, establishing ongoing measures will encourage a more focused and effective discussion and coordination of risks. It is important that we continue to build on understanding the most appropriate ways to render these measures and integrate them through basic leadership for general well-being. Improving ongoing speculations of inevitable events of illness is a significant specialized achievement, as study and concerted efforts are anticipated to create a superior understanding of how to impart these outcomes to leaders of general well-being and integrate irresistible disease standards into general well-being practice. A big challenge for the epidemiological and biostatistical networks of experts and scholastics is to find out how to turn this data storm into evidence that educates basic leadership on enhancing well-being and preventing individual and population-level disease. The community effort represented in this paper offers a framework for establishing consistent objectives by and by and reflects clear outcomes from this move to integrate current information technology resources with fundamental leadership for general well-being.

REFERENCES

[1] Senthilkumar Mohan et al, “Efficient Heart Disease prediction using hybrid machine learning techniques”, IEEE access, 7, 2019.

[2] C. Ordonez, “Association rule discovery with the train and test approach for heart disease prediction”, IEEE transactions on Information Technology in biomedicine, 10(2), 2006.

[3] Bio Jin et al, “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling”, IEEE access, 6, 2018.

[4] Li Na Pu et al, “Investigation on Cardiovascular Risk Prediction Using Genetic Information”, IEEE transactions on Information Technology in biomedicine, 16(5), 2012.

[5] Chrysa D. Papadaniil et al, “Efficient Heart Sound Segmentation and Extraction Using Ensemble Empirical Mode Decomposition and Kurtosis Features”, IEEE journal of biomedical and Health informatics, 18(4), 2014.

[6] Li, H. L., Lin, S. W., & Hwang, Y. T. (2019). Using nursing information and data mining to explore the factors that predict pressure injuries for patients at the end of life. CIN: Computers, Informatics, Nursing, 37(3), 133-141.

[7] Oladele, T. O., Olarinoye, D. R., & Adebisi, S. S. (2018). PREDICTION OF SKIN DISEASE USING DECISION TREE AND ARTIFICIAL NEURAL NETWORK (ANN). Annals. Computer Science Series, 16(1).

[8] Anurag Kumar Verma et al, “Comparison of skin disease prediction by feature selection using ensemble data mining techniques”, Informatics in Medicine unlocked, 2019.

[9] Han Wu et al, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine unlocked, 10, 2018.

[10] Nfonguourain Mougnutou Remy, “The prediction of good physicians for prospective diagnosis using data mining”, Informatics in Medicine unlocked, 12, 2018.

[11] J. C. Madden et al, “In silicoprediction of skin metabolism and its implication in toxicity assessment”, Computational Toxicology, 3, 2017.

[12] Shuai Ding, “Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model”, Information Sciences, 499, 2019.

[13] M. Sharma et al, “Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques”, Elsevier Masson, 38(6), 2017.

14] Changsheng Zhu et al, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques”, Informatics in Medicine Unlocked, 2019.

[15] Ahmed, S. R. A., Al Barazanchi, I., Mhana, A., & Abdulshaheed, H. R. (2019). Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set. Periodicals of Engineering and Natural Sciences, 7(2), 438-447.

[16] Behadili, S. F., Abd, M. S., Mohammed, I. K., & Al-SAYYID, M. M. (2019). Breast Cancer Decisive Parameters for Iraqi Women via Data Mining Techniques. Journal of Contemporary Medical Sciences, 5(2).

[17] Johnston, W. L., Catton, C. N., & Swallow, C. J. (2019). Unbiased data mining identifies cell cycle transcripts that predict non-indolent Gleason score 7 prostate cancer. BMC urology, 19(1), 4.

[18] AbouElNadar, N. A., & Saad, A. A. (2019, July). Towards a Better Model for Predicting Cancer Recurrence in Breast Cancer Patients. In Intelligent Computing-Proceedings of the Computing Conference (pp. 887-899). Springer, Cham.

19] Bratislav Predic et al, “ Data mining based tool for early prediction of possible fruit pathogen infection”, Computers and Electronics in Agriculture, 154, 2018.

[20] Arif Khan et al, “Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes”, Experts systems with application, 136, 2019.