IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Disease Prediction Using Machine Learning

Rakhi Singh, Student, School of Computer Science and Engineering, Galgotias University, Greater Noida, India

Tushar Singh, Student, School of Computer Science and Engineering, Galgotias University, Greater Noida, India

Mr. Ajay Shankar, , School of Computer Science and Engineering, Galgotias University, Greater Noida, India

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases va rious conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt man agement of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to p redict heart disease. This research paper presents various attributes rel ated to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dat aset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of thes e 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is ach ieved with K-nearest neighbor.

LITERATURE SURVEY

A. Common DiseasesDahiwade et al. [9] proposed a ML based system that pre-dicts common diseases. The symptoms dataset was importedfrom the UCI ML depository, where it contained symptomsof many common diseases. The system used CNN and KNNas classification techniques to achieve multiple diseases pre-diction. Moreover, the proposed solution was supplementedwith more information that concerned the living habits of thetested patient, which proved to be helpful in understanding thelevel of risk attached to the predicted disease. Dahiwade et al.[9] compared the results between KNN and CNN algorithmin terms of processing time and accuracy. The accuracyand processing time of CNN were 84.5% and 11.1 seconds,respectively. The statistics proved that KNN algorithm is underperforming compared to CNN algorithm. In light of thisstudy, the findings of Chen et al. [10] also agreed that CNNoutperformed typical supervised algorithms such as KNN, NB,and DT. The authors concluded that the proposed model scoredhigher in terms of accuracy, which is explained by the capability of the model to detect complex nonlinear relationships inthe feature space. Moreover, CNN detects features with highimportance that renders better description of the disease, whichenables it to accurately predict diseases with high complexity[9], [10]. This conclusion is well supported and backed withempirical observations and statistical arguments. Nonetheless,the presented models lacked details, for instance, NeuralNetworks parameters such as network size, architecture type,learning rate and back propagation algorithm, etc. In addition,the analysis of the performances is only evaluated in terms ofaccuracy, which debunks the validity of the presented findings[9]. Moreover, the authors did not take into considerationthe bias problem that is faced by the tested algorithms [9],[10]. In illustration, the incorporation of more feature variablescould immensely ameliorate the performance metrics of underperformed algorithms

[11].B. Kidney DiseasesSerek et al. [12] planned a comparative study of classifiersperformance for Chronic Kidney disease (CKD) detectionusing The Kidney Function Test (KFT) dataset. In this study,the classifiers used are KNN, NB, and RF classifier; theirperformance is examined in terms of F-measure, precision,and accuracy. As per analysis, RF scored better in phrases ofF-measure and accuracy, while NB yielded better precision. Inconsideration of this study, Vijayarani [13] aimed to detect kid-ney diseases using SVM and NB. The classifiers were used toidentify four types of kidney diseases namely Acute NephriticSyndrome, Acute Renal Failure, Chronic Glomerulonephritis,and CKD. Additionally, the research was focused on deter-mining the better performing classification algorithm basedon the accuracy and execution time. From the results, SVMconsiderably achieved higher accuracy than NB, which makesit the better performing algorithm. However, NB classifieddata with minimum execution time. Other several empiricalstudies also focused on locating CKD; Charleonnan et al. [14]and Kotturu et al. [15] concluded that the SVM classifier isthe most adequate for kidney diseases because it deals wellwith semi-structured and unstructured data. Such flexibilityallowed SVM to handle larger features spaces, which resultedin acquiring high accuracy when detecting complex kidneydiseases. Although supported by findings, the conclusion isweakened by prior suggestion that different hyper-parameterswere not experimented when evaluating the performances ofML algorithms. According to Uddin [3] the exploration of thehyper-parameter space can generate different accuracy resultsand render better performances for ML algorithms.C. Heart DiseasesMarimuthu et al. [16] aimed to predict heart diseasesusing supervised ML techniques. The authors structured theattributes of data as gender, age, chest pain, gender, target andslope [16]. The applied ML algorithms that were deployedare DT, KNN, LR and NB. As per analysis, the LR algorithmgave a high accuracy of 86.89%, which deemed to be the mosteffective compared to the other mentioned algorithms. In 2018,Dwivedi [17] attempted to add more precision to the predictionof heart diseases by accounting for additional parameters suchas Resting blood pressure, Serum Cholesterol in mg/dl, andMaximum Heart Rate achieved. The used dataset was importedfrom the UCI ML laboratory; it was comprised with 120samples that were heart disease positive, and 150 samples thatwere heart disease negative. Dwivedi attempted to evaluatethe performance of Artificial Neural Networks (ANN), SVM,KNN, NB, LR and Classification Tree. At the appliance oftenfold cross validation, the results showed that LR has thehighest classification accuracy and sensitivity, which showshigh dependability at detecting heart diseases [17]. This con-clusion is strengthened by the findings of Polaraju [18] andVahid et al. [19], where the Logistic Regression outperformedother techniques such as ANN, SVM, and Adaboost. Thestudies excelled in conducting an extensive analysis on the MLmodels. For instance, various hyper-parameters were tested ateach ML algorithm to converge to the best possible accuracyand precision values. Despite that advantage, the small sizeof the imported datasets constraints the learning models fromtargeting diseases with higher accuracy and precision

MODULE DESCRIPTION

The existing system predicts the chronic diseases which are for a particular region and for the particularcommunity. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithmis used for Disease risk prediction. For S type data, the system is using Machine Learning algorithm i.e K-nearest Neighbors, Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%. In theexisting paper, they streamline machine learning algorithms for the effective prediction of chronic diseaseoutbreak in disease-frequent communities. They experiment with the modified prediction models over reallifehospital data collected from central China. They propose a convolutional neural network-based multimodaldisease risk prediction(CNN-MDRP) algorithm using structured and unstructured data from the hospital.

Most of the chronic diseases are predicted by our system. It accepts the structured type of data as input to themachine learning model. This system is used by end-users i.e. patients/any user. In this system, the user willenter all the symptoms from which he or she is suffering. These symptoms then will be given to the machinelearning model to predict the disease. Algorithms are then applied to which gives the best accuracy. ThenSystem will predict disease on the basis of symptoms. This system uses Machine Learning Technology. NaïveBayes algorithm is used for predicting the disease by using symptoms, for classification KNN algorithm isused, Logistic regression is used for extracting features which are having most impact value, the Decision treeis used to divide the big dataset into smaller parts. The final output of this system will be the disease predicted.

To calculate performance evaluation in the experiment, first, we denote TP, TN, Fp and FNias

truepositive(the number of results correctly predicted as required), true negative (the number of results notrequired), false positive (the number of results incorrectly predicted as required), false negative(the number ofresults incorrectly predicted as not required)respectively

KNN K Nearest Neighbour (KNN) could be terribly easy, simple to grasp, versatile and one amongst theuppermost machine learning algorithms. In the Healthcare System, the user will predict the disease. In thissystem, the user can predict whether the disease will detect or not. In the proposed system, classifying diseasein various classes that shows which disease will happen on the basis of symptoms. KNN rule used for eachclassification and regression issue. KNN algorithm is based on feature similarity approach. It is the bestchoice for addressing some of the classification related tasks. K-nearest neighbor classifier algorithm is topredict the target label of a new instance by defining the nearest neighbor class. The closest class will beidentified using distance measures like Euclidean distance. If K = 1, then the case is just assigned to thecategory of its nearest neighbor.The value of 'k' has to be specified by the user and the best choice depends on the data. Thelarger value of 'k' reduces the noise on the classification. If the new feature i.e in our case symptom has toclassify, then the distance is calculated and then the class of feature is selected which is nearest to the newerinstance. In the instance of categorical variables, the Hamming distance must be used. It conjointly brings upthe difficulty of standardization of the numerical variables between zero and one once there's a combination ofnumerical and categorical variables within the dataset

Naive Bayes is an easy however amazingly powerful rule for prognosticative modeling. The independenceassumption that allows decomposing joint likelihood into a product of marginal likelihoods is called as 'naive'.This simplified Bayesian classifier is called as naive Bayes. The Naive Bayes classifier assumes the presenceof a particular feature in a class is unrelated to the presence of any other feature. It is very easy to build anduseful for large datasets. Naive Bayes is a supervised learning model. Bayes theoremmprovides some way ofcalculative posterior chance P(b|a) from P(b), P(a) and P(a|b).

A decision tree is a structure that can be used to divide up a large collection of records into successfullysmaller sets of records by applying a sequence of simple decision tree. With each successive division, themembers of the resulting sets become more and more similar to each other. A decision tree model consists ofa set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutuallyexclusive) groups with respect to a particular target. The target variable is usually categorical and the decisiontree is used either to:

- Calculate the probability that a given record belong to each of the category and,
- To classify the record by assigning it to the most likely class (or category). In this disease predictionsystem, decision tree divides the symptoms as per its category and reduces the dataset difficulty .

RESULTS

Comparison of accuracy of algorithm.Decision Tree 84.5%Random Forest 98.95%Naïve Bayes 89.4%SVM 96.49%KNN 71.28% We found that the Support Vector Machine (SVM) algorithm is widely used (in 30 studies)followed by the Naïve Bayes algorithm (in 24 studies). However, the Random Forest algorithm showedrelatively high accuracy. In the 40 studies in which it was used, RF showed the highest accuracy of 98.95%.This was followed by SVM which included 96% of the accuracy considered.

CONCLUSION

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. Thissystem takes the symptoms of the user from which he or she suffers as input and generates final output as aprediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor wassuccessfully implemented using the grails framework. This system gives a user-friendly environment and easyto use.As the system is based on the web application, the user can use this system from anywhere and at any time. Inconclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of thehospital data.This systematic review aims to determine the performance, limitations, and future use of Software in healthcare. Findings may help inform future developers of Disease Predictability Software and promotepersonalized patient care. The program predicts Patient Diseases. Disease Prediction is done through UserSymbols.In this System Decision tree, Unplanned Forest, the Naïve Bayes Algorithm is used to predict diseases. Forthe data format, the system uses the Machine Learning algorithm Process Data on Database Data namely,Random Forest, Decision Tree, Naive Bayes. System accuracy reaches 98.3%. machine learning skills aredesigned to successfully predict outbreaks.

REFERENCES

1. Dr.C K Gomathy, Article: A Semantic Quality of Web Service Information Retrieval Techniques UsingBin Rank A Cloud Monitoring Framework Perform in Web Services, International Journal of ScientificResearch in Computer Science Engineering and Information Technology IJSRCSEIT | Volume 3 | Issue 5 |ISSN : 2456-3307,May-2018
2. Dr.C K Gomathy, Article: Supply chain-Impact of importance and Technology in Software ReleaseManagement, International Journal of Scientific Research in Computer Science Engineering andInformation Technology (IJSRCSEIT) Volume 3 | Issue 6 | ISSN : 2456-3307, P.No:1-4, July-2018
3. Dr.C K Gomathy, Article: A Scheme of ADHOC Communication using Mobile Device Networks,International Journal of Emerging technologies and Innovative Research (JETIR) Volume 5 | Issue 11 |ISSN : 2349-5162, P.No:320-326, Nov-2018
4. Dr.C K Gomathy, Article: A Study on the recent Advancements in Online Surveying , InternationalJournal of Emerging technologies and Innovative Research (JETIR) Volume 5 | Issue 11 | ISSN : 2349-5162, P.No:327-331, Nov-2018
5. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, no G. Escobar, "Big data for health care: usinganalytics to identify and treat high-risk and high-risk patients, Health, vol. 33, no. 7, pages 1123–1131,2014.
6. K.R. Lakshmi, Y. Nagesh and Mr. Veera Krishna, "Comparison of performance of the three data minesways to predict survival of kidney disease", International Journal of Engineering Development &Technology, March 2014
7. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction ofheart disease using machine learning," in 2018 Second InternationalConference on Electronics, Communication and Aerospace Technology(ICECA), 2018, pp. 1275–1278.
8. Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatolog-ical disorders through image-processing and machine learning," in 2017International Conference on Intelligent Sustainable Systems (ICISS),2017, pp. 1047–1051.
9. S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparingdifferent supervised machine learning algorithms for disease prediction,"BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019
10. R. Katarya and P. Srinivas, "Predicting heart disease at early stagesusing machine learning: A survey," in 2020 International Conferenceon Electronics and Sustainable Communication Systems (ICESC), 2020,pp. 302–305.
11. P. S. Kohli and S. Arora, "Application of machine learning in diseaseprediction," in 2018 4th International Conference on Computing Com-munication and Automation (ICCCA), 2018, pp. 1–4.
12. M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra,"A proposed model for lifestyle disease prediction using support vectormachine," in 2018 9th International Conference on Computing, Com-munication and Networking Technologies (ICCCNT), 2018, pp. 1–6.
13. F. Q. Yuan, "Critical issues of applying machine learning to conditionmonitoring for failure diagnosis," in 2016 IEEE International Confer-ence on Industrial Engineering and Engineering Management (IEEM),2016, pp. 1903–1907.
14. S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning ma-chine (elm) technique for heart disease diagnosis," in 2015 IEEE CanadaInternational Humanitarian Technology Conference (IHTC2015), 2015

D. Dahiwade, G. Patle, and E. Meshram, "Designing disease predictionmodel using machine learning approach," Proceedings of the 3rd Inter-national Conference on Computing Methodologies and Communication,ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.
15. S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, "Disease Predictionby Machine Learning from Healthcare Communities," InternationalJournal of Scientific Research in Science and Technology, pp. 29–35,2019.
16. R. Saravanan and P. Sujatha, "A state of art techniques on machinelearning algorithms: A perspective of supervised learning approachesin data classification," in 2018 Second International Conference onIntelligent Computing and Control Systems (ICICCS), 2018, pp. 945–949.
17. Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronickidney disease dataset by applying machine learning methods," in 2018IEEE 12th International Conference on Application of Information andCommunication Technologies (AICT), 2018, pp. 1–4.
18. V. S and D. S, "Data Mining Classification Algorithms for Kidney Dis-ease Prediction," International Journal on Cybernetics & Informatics,vol. 4, no. 4, pp. 13–25, 2015.
19. A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit,S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronickidney disease using machine learning techniques," 2016 Managementand Innovation Technology International Conference, MITiCON 2016,pp.

MIT80–MIT83, 2017.[15] P. Kotturu, V. V. Sasank, G. Supriya, C. S. Manoj, and M. V. Mahesh-warredy, "Prediction of chronic kidney disease using machine learningtechniques," International Journal of Advanced Science and Technology,vol. 28, no. 16, pp. 1436–1443, 2019.

20. M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra,"A Review on Heart Disease Prediction using Machine Learning andData Analytics Approach," International Journal of Computer Applica-tions, vol. 181, no. 18, pp. 20–25, 2018.