International Journal of IT & Knowledge Management (ISSN: 0973-4414)
Volume 15 • Issue 1 pp. 23-28 June 2021 - Dec 2021 www.csjournals.com

Utilization of Machine Learning Models for Disease Prediction

Jaswinder Singh[1], Neha Bhadu[2]
Department of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India
jaswinder_singh_2k@rediffmail.com, nehabhadu100@gmail.com

ABSTRACT

Machine learning in health sector has a great ability to find the hidden pattern in medical datasets. These patterns aids in diagnosis and prognosis of different diseases. The raw medical data is generally unstructured, distributed, large and very complex. Therefore, it becomes manually impossible for the healthcare professional to process the data. Examining this data requires lots of efforts, time and money. Identification of the disease is very important step for its treatment. Machine learning has the ability to learn patterns from data and then draw inferences to make predications.This study delivers the fundamental knowledge of ML techniques along with its approaches. Also it provides the extensive review based on utilization of machine learning models to predict different diseases.

Keywords: *Artificial Intelligence (AI), Machine Learning (ML), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Neural Network (NN)*

I. Introduction

In present connected globe data is increasing exponentially across different fields. As a resultof which technologies that interact with data to extract meaning from the information are rising and machine learning (ML) is one amongst them. ML is the fast emerging technological field that lies at the junction of computer science and statistics, and at the centre of artificial intelligence (AI) and data science. It addresses the issue of constructing such computers which can involuntarily advance themselves all the way through experience. Presently, there are numerous sectors where ML is deeply functional. In particular, one sector where ML is practiced at large is healthcare realm. Healthcare is one of the rapidly rising fields that can extensively benefit from the growing number of statistics and its accessibility. One of the main uses of ML in the medical field is to detect and analyze diseases that are difficult to identify. It is being broadly utilized in timely drug discovery process, medicinal imaging analysis, custom-made medication, smart healthiness files, scientific testing and investigation, crowd-sourced information gathering, outburst forecast and several others. ML is recognized as having key technological application in transforming the healthcare sector benefiting both patients and clinicians. The prime objectives of this work are mentioned below:

- To acquire fundamental knowledge of ML and its Techniques.
- To provide with an extensive review of the ML predictive models employed in predicting different diseases.
- To discover the opportunities and challenges that subsists when ML techniques are employed in prediction of diseases.

The remaining paper is arranged as follows:- Section II provides a short introduction of ML along with its types. Section III explains different approaches to ML. Section IV gives the systematic review of the literature studied for predicting different diseases. Section V and VI presents the benefits and challenges found in utilizing ML in disease prediction. The last section presents the conclusion followed by the references.

II. Machine Learning

ML is a subfield of AI providing machines with the capability to involuntarily discover from data, enhance performance from practices, and forecast things with no explicit programming. It is the learning of computer programs which repeatedly enhance themselves through training [1]. The ML model acquires knowledge from past data, develops the predictive systems, and as soon as it collects new information, calculates the outcome for it. The correctness of estimated outcome depends on the volume of data, as large volume of data facilitates in building more robust and reliable model which can calculate the outcome with more accuracy. Generally, ML is divided into 4 types as shown in Table 1.1 beneath [2]:

Table 1.1 Different ML Types along with their Example

Category of Learning	Model Development and Approach	Examples
Supervised Learning	• Models and Algorithms learn through labeled data • Driven by Task	Regression, Classification etc.
Unsupervised Learning	• Models and Algorithms learn through unlabeled data • Driven by Data	Association, Principal Component Analysis Clustering, etc.
Semi-supervised Learning	• Collective data i.e. a combination of labeled and unlabeled data is used for model's development	Clustering, Classification etc.
Reinforcement Learning	• Models are built on the basis of prize or punishment • Driven by Environment	Control, Q-learning, Classification etc.

III. Machine Learning Approaches

ML algorithm develops a statistical model on the basis of training data. Different algorithms that promise to deliver excellent diagnostics in healthcare are mentioned below [3]:

1. K-Nearest Neighbour (KNN): For classification issues, pattern recognition, and regression, it is a straightforward model. Euclidean distance between data points is used to find neighbours among the data. For classification and regression issues, it is used. In order to locate the new case for a related category, the value of k will find all of the existing feature cases that are comparable to the new case and surround all cases. As a result, the K value is crucial and must be carefully set since if it is too little, the system may become overfit.

2. Support Vector Machine (SVM): This form of unsupervised ML technology is widely applied in the closest neighbor based clustering method. According on how similar they are to one another, the data can be divided into k clusters. K is an integer, and for the algorithm to work, we need to know what it is. The most popular clustering algorithm is K-mean, which may select the best cluster for brand-new data based on the majority of the distance.

3. Naive Bayes (NB): It is a mathematical and probabilistic approach based on a classification algorithm. It is a common technique in machine learning applications because it makes it possible for all features to have an equal impact on the conclusion. This simplicity is matched by computational efficiency, which makes the NB technique intriguing and suitable for various applications.

4. Decision Tree (DT): It is a mathematical and probabilistic approach based on a classification algorithm. It is a common technique in machine learning applications because it makes it possible for all features to have an equal impact on the conclusion. This simplicity is matched by computational efficiency, which makes the NB technique intriguing and suitable for various applications.

5. Random Forest (RF): It is an ensemble technique that can also be used to make predictions based on your nearest neighbours. The fundamental tenet of ensemble techniques is that a collection of models will combine to create a potent model. A standard ML method decision tree in ensemble terms is included with the random forest.

6. Logistic Regression (LR): It is a supervised machine learning technique used to address a binary classification problem. Binary classification is modelled mathematically using the logistic function, and there are various more complex extensions for LR. In its simplest form, LR is a regression model that estimates the likelihood that a given data point or entry will fall into a particular class using the regression model.

7. Ensemble: This method combines many algorithm types to improve the model's performance and prognostic capabilities. Simple models sometimes fail to produce accurate findings because of strong bias or excessive variation. In that circumstance, robust models are constructed using ensemble approaches. The trio of bagging, boosting, and stacking ensembles is the most popular.

IV. Literature Review

Researchers have applied different machine learning algorithms and techniques to build a predictive model for theprediction of diseases. Some of the work done by them is mentioned below:

R. Bhardwaj *et al.* **(2017) in [4],** aimed to study the significances of ML in disease prediction, diagnostics and ML-based health care applications. It also outlined various industries using machine learning initiatives in the healthcare area. This study has found that ML has the potential to equally help both providers and patients in the terms of lower costs and better care respectively.

Thirunavukkarasu *et al.* **(2018) in [5],** presented a model that predicts liver diseases using diverse classification techniques such as LR, SVM and KNN. The accuracies of LR and KNN were equal but the sensitivity result of LR was highest. It was found from the performance results that LR was best in prediction of liver diseases.

Ishita Kapoor and Anju Mishra in 2018 [6], evaluated the neural network which helped in detection of alopecia areata in human beings and also calculated the accuracy. The proposed system used feed-forward ANN and back-propagation algorithm to do classification of patients with and without alopecia. An accuracy of 91% was achieved which is enough for clinical experts to take good quality of decisions.

A. Mir and S. N. Dhage (2018) in [7], showed that a classifier system used WEKA tool forprediction of diabetes by making use of NB, SVM, RF and Simple CART algorithms. Afterwards these classifiers were compared on the basis of training time, accuracy value andtesting time. It was found out that SVM's accuracy was highest and it outperformed other classifiers inpredicting diabetes.

P. S. Kohli and S. Arora (2018) in [8], presented several classification algorithms using datasets on three different diseases i.e. heart, diabetes and breast cancer accessible in UCI datasets for disease prediction. The datasets were selected using backward modelling technique that utilized the score of p-value analysis. Accuracy of prediction in case of heart disease was found to be 87.1% utilizing LR algorithm, in case of diabetes by utilizing SVM algorithm the accuracy was found to be 85.71% and in case of breast cancer by using AdaBoost algorithm the accuracy reached 98.5%.

Mehrbakhsh Nilashi *et al.* **(2019) in [9],** developed a model for diagnosing hepatitis disease by utilizing advantageous ensemble learning. "Non-linear Iterative Partial Least Squares Technique" was used for performing dimensionality reduction of the data. Self organising maps were used for grouping. For the prediction process ensembles of "NeuroFuzzy Inference System" were utilized. For collecting the most significant characteristics of experimental data, DT was used. The study concluded that the proposed system outperformed NN, ANFIS, KNN and SVM.

D. Dahiwade *et al.* **(2019) in [10],** proposed a general model for disease prediction which was developed on the basis of the patient's symptoms. KNN and CNN algorithms were utilized for the accurate disease prediction.Disease prediction models accurateness utilizing CNN was found to be 84.50% that was greater than that of KNN. Furthermore space and time complexity was foundhigher in KNN as compared to CNN.

B. Davi *et al.* **(2019) in [11],** projected a ML technique for prediction of severity of dengue fever. Human's genome information was the bases of this technique. The model used SVM algorithm to figure out the most favourable loci's classification division. Afterwards, ANN was utilized to categorize patients with severe dengue fever. ANN model produced median values with precision larger than 86%. The values for sensitivity and specificity were found to be over 98% and 51% respectively.

G. Çınarer and B. G. Emiroğlu (2019) in [12], studied different classification methods regarding tumours. These methods were employed for categorizing the characteristics of the brain's image into multifical, n/a, multi-centric and glimatosis. During the classification procedure the images were examined for their arithmetical properties and afterwards data was segmented into different groups. KVM, LDA, SVM and RF

classifiers were utilized for testing the data. SVM outperformed other classifiers and acquired an accuracy of 90%.

Q. Atallah and A. Al-Mousa (2019) in [13], presented a ML ensemble technique thatcombined multiple ML techniques like KNN, RF, LR and hard voting ensemble classifier for heart disease prediction.Ensemble model's correctness was found to be 90%, which exceeds the accuracy provided by each of the individual classifier.

C. Jain *et al.* (2019) in [14], compared several progressions powered by ML and data mining, the future of ML in the healthcare and the role of ambient intelligence system. The study found that ML can support in the detection, diagnosis and prevention of diseases.

K. R. Dalal (2020) in [15], addressed applicability of ML in medical field and put forward several difficulties that health sector possibly will tackle at various steps while executing ML for a variety of purposes. The study concluded that identifying and rectifying different problems related to data is of utmost importance to fully exercise the capabilities of ML.

N.K. Kumar *et al.* **(2020) in [16],** utilized several ML algorithms in the identification of cardio vascular disease (CVD). The classifiers used were RF, LR, DT, KNN and SVM. The RF classifier attained an accurateness of 85.71% and value of ROC-AUC was found to be 0.8675. The study concluded that RF classifier performed best in comparison to all other classifiers taken into consideration for categorising patients having CVD.

Luca Brunese *et al.* **(2020) in [17],** developed a model for the identification of COVID-19 automatically by examining the medicinal images. KNN algorithm was utilized for building the model. While differentiating COVID-19 positive patients and the patients with other pulmonary ailments having identical signs, the model attained the recall and an average precision of 0.965. The experimental results demonstrated the goodness of the model proposed.

H. E. Hamdaou *et al.* **(2020) in [18],** developed a medical support model for making good choices while identifying heart disease. For that purpose the study utilized NB, KNN, SVM, RF and DT algorithmsThe results of cross-validation as well as train test split showed that NB performed better and their accurateness was found to be 82.17% and 84.28% in the respective tests. The study also found that the accuracy of each classifier declines after employing cross-validation method.

Islam *et al.* (2020) in [19], compared five supervised ML techniques namely SVM, RF, KNN, LR and ANN. Breast cancer dataset from Wisconsin was used for carrying out the study and it was collected from UCI repository.The experimental results demonstrated that ANN attained maximum F1-score, accuracy and precision among all classifiers and there values were 0.9890, 98.57% and 97.82% respectively.

Harimoorthy, K. and Thangavelu, M. (2020) in [20], proposed a structure to predict multiple diseases namely chronic kidney disease, diabetes and heart disease.Chi square technique was utilized to identify the important characteristics. The experiments were carried out with "SVM-Linear, SVMPolynomial, Improved SVM-Radial bias kernel, RF and DT" techniques. The results demonstrated that improved SVM-Radial bias kernel showed better accuracy when compared with other techniques. For CKD, heart disease and diabetes it showed an accuracy of 98.3%, 89.9% and 98.7% correspondingly.

Jianhong Kang *et al.* **(2021) in [21],** developed a model dependent on the medical data for the identification of COVID-19 patients. The data was obtained from "Tumor Centre of Union Hospital, China". Utilizing NN the system was built by applying TensorFlow consequently. The experimental results showed good performance of model in prediction. The value for AUC was found to be 0.953.

Chaubey, G. *et al.* **(2020) in [22],** compared most commonly used three ML techniques namely KNN, DT and LR to examine their performance in measures of accurateness. The study highlighted the utilization of these ML classifiers as equipment for the prediction of thyroid. The results showed that KNN outperformed other classifiers in predicting thyroid.

V. Benefits of Disease Prediction using ML

Early stage diagnosis increases the chances of curing the disease and minimizes the death rate. In case of communicable diseases early diagnosis of a disease will help in stopping the further spread of the disease by taking preventive measures. The model can also be utilized by the doctors for taking second opinion while

diagnosing diseases and also it will help them diagnose more patients is less time. It will also be of great help in diagnosing diseases for which the testing resources are limited. Therefore employing ML in disease prediction will reduce time of practitioner to diagnose the disease so that practitioner could check more patients.It will also reduce effort of practitioner in regard to treat same kind of disease.It will make medical history of a patient available in emergency situation.Also it will help in saving lives of the patients by diagnosing early.

VI. Challenges in Disease Prediction using ML

The biggest challenge in disease prediction using ML is being able to obtain a patient data set with the quality and sample size necessary to train a recent ML model [23].Data collection, sharing and distribution are not easy as patient data is protected by strict confidentiality and safety rules. The data pipeline should be continually revised to reveal the actual world, not just a single snapshot. It is required to optimize the data and make it consistent prior using it to enhance the technical data. Learning from data to answer the question of causality is a difficult task [24].Truly complete data on cost and volume are often impractical. Learning from incomplete or missing data is a difficult task. Predictions obtained using machine learning algorithms are only accurate if there is no malicious data in the training dataset. The ML algorithm can mirror individual biases during the decision-making process.

Conclusion

ML plays an important role in accomplishing the timely prediction of diseases. This study provided a systematic review of diverse ML algorithms for predicting diseases and standard data sets have been utilized for carrying out the research. A table has been formulated representing the research gaps found in the literature reviewed. It was found that the accurateness of the same algorithm may vary from one dataset to another because numerous essential components impacts the model's accurateness and performance such as datasets taken, feature selection, dimentionality reduction techniques used and the number of features taken to carry out the research. Another important significance drawn from this study is that the model's accurateness and performance can be augmented by utilizing different ML algorithms to produce one ensemble model.

REFERENCES

[1] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
[2] I. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*, vol. 2, no. 3, 2021. Available: 10.1007/s42979-021- 00592-x.
[3] Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms forDiagnosing Diseases", Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 10-19, 2021. Available: 10.38094/jastt20179.
[4] R. Bhardwaj, A. Nambiar and D. Dutta, "A Study of Machine Learningin Healthcare", *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2017. Available:10.1109/compsac.2017.164.
[5] K. Thirunavukkarasu, A. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms", *2018 4th International Conference on Computing CommunicationandAutomation(ICCCA)*,2018.Available:10.1109/ccaa.2018.8777655.
[6] I. Kapoor and A. Mishra, "Automated Classification Method for Early Diagnosis of Alopecia Using Machine Learning", *Procedia Computer Science*, vol. 132, pp. 437-443, 2018. Available:10.1016/j.procs.2018.05.157.
[7] A. Mir and S. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare", *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018. Available:10.1109/iccubea.2018.8697439.
[8] P. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction", *2018 4thInternationalConferenceonComputingCommunicationandAutomation(ICCCA)*,2018. Available:10.1109/ccaa.2018.8777449.
[9] M.Nilashi,H.Ahmadi,L.Shahmoradi,O.IbrahimandE.Akbari,"Apredictivemethod for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique", *Journal of Infection and Public Health*, vol. 12, no. 1, pp. 13-20, 2019. Available: 10.1016/j.jiph.2018.09.009.
[10] 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)*, pp. 1211-1215, 2019. Available: 10.1109/iccmc.2019.8819782.

[11] C. Davi et al., "Severe Dengue Prognosis Using Human Genome Data and Machine Learning", *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2861-2868, 2019. Available:10.1109/tbme.2019.2897285.
[12] G. Cinarer and B. Emiroglu, "Classification of Brain Tumors by Machine Learning Algorithms", *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2019. Available:10.1109/ismsit.2019.8932878.
[13] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method", *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 2019. Available:10.1109/ictcs.2019.8923053.
[14] D. Jain, B. Kadecha and S. Iyer, "A Comparative Study of Machine Learning Techniques in Healthcare,"*2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2019, pp.455-460.
[15] K. Dalal, "Analysing the Implementation of Machine Learning in Healthcare", *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 133-137, 2020. Available:10.1109/icesc48915.2020.9156061.
[16] N. Kumar, G. Sindhu, D. Prashanthi and A. Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 15-21, 2020. Available:10.1109/icaccs48705.2020.9074183.
[17] L. Brunese, F. Martinelli, F. Mercaldo and A. Santone, "Machine learning for coronavirus covid-19 detection from chest x-rays", *Procedia Computer Science*, vol. 176,pp. 2212-2221, 2020. Available:10.1016/j.procs.2020.09.258.
[18] H. Hamdaoui, S. Boujraf, N. Chaoui and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques", *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* , pp. 1-5, 2020. Available:10.1109/atsip49331.2020.9231760.
[19] M. Islam, M. Haque, H. Iqbal, M. Hasan, M. Hasan and M. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", *SN Computer Science*, vol. 1, no. 5, 2020. Available:10.1007/s42979-020-00305-w.
[20] Harimoorthy,K.,Thangavelu,M. ”Multi-disease prediction model using improvedSVM- radialbias technique in healthcare monitoring system. *J Ambient Intell Human Comput* **12,** 3715–3723 (2020).https://doi.org/10.1007/s12652-019-01652-0.
[21] J. Kang, T. Chen, H. Luo, Y. Luo, G. Du and M. Jiming-Yang, "Machine learning predictive model for severe COVID-19", *Infection, Genetics and Evolution*, vol. 90, p. 104737, 2021. Available:10.1016/j.meegid.2021.104737.
[22] G. Chaubey, D. Bisen, S. Arjaria and V. Yadav, "Thyroid Disease Prediction Using Machine Learning Approaches", *National Academy Science Letters*, vol. 44, no. 3, pp. 233- 238, 2020. Available:10.1007/s40009-020-00979-z.
[23] R. Miotto, F. Wang, S. Wang, X. Jiang and J. Dudley, "Deep learning for healthcare: review, opportunities and challenges", *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236- 1246, 2017. Available:10.1093/bib/bbx044.
[24] M. Ghassemi, T. Naumann, P. Schulam, A. Beam, I. Chen and R. Ranganath, "Practical guidanceonartificialintelligenceforhealth-caredata",*TheLancetDigitalHealth*,vol.1,no. 4, pp. e157-e159, 2019. Available:10.1016/s2589-7500(19)30084-6.