Diabetes Disease Prediction Using Machine Learning

Preetha S¹, Chandan N², Darshan N K³, Gowrav P B⁴
¹,²,³,⁴ Department Of ISE, B.M.S. College Of Engineering, VTU, Bengaluru, India

Abstract— Data mining is known as the process of sorting through a large number of data sets to create relationships and to find patterns for solving a given problem through data analysis. Classification, Association rules, Prediction, Clustering, and Sequential patterns are some of the most important and common data mining techniques. Data mining methods are used in a wide variety of applications, also seen in the healthcare sector. Data analysis in the healthcare industry plays a significant role in the detection of diseases. A variety of tests would be expected from the patient to diagnose a certain disease. However, using the data mining technique, the number of tests can be minimized. These tests play a critical role in time and results. The data mining methodology has benefits and drawbacks. This paper analyses how data mining methods are used to identify various types of diseases. This paper surveyed research articles that focused primarily on predicting, diabetes, heart disease and skin cancer.

Keywords— Disease, Diabetes, Prediction, Naïve Bayes, KNN

I. INTRODUCTION

Machine learning can be divided into four forms, Un-supervised learning; Supervised learning; Semi-supervised learning; and Reinforcement learning. Un-Supervised Learning are often unlabelled data that may be inadvertently missing labels or deliberately missing labels. Former specimens are usually labelled initially; the label can be eliminated and the question introduced as a relationship or association analysis between samples. Explanations for these are two fold, First of all, the personnel can fail to insert the mark by accident. Second, it is likely that the data owner company does not feel that the data are valuable to ensure that data collection management is ignored.

Supervised Learning: In machine learning, due to its use in regression and classification, a broader community of researchers focuses on supervised learning. The five general steps for the supervised learning model are

• Data collection of training and testing datasets
• Extraction of features;
• Selection of machine learning algorithms;
• Model creation using the chosen algorithm; and
• Algorithm validation and comparison with other algorithms.

Semi-Supervised Learning: Semi-supervised learning is relatively young compared to unsupervised and supervised learning. Chances are that there are a lot of unmarked data, but only a few classified data are available. This is the reason for the semi-supervised term, as it lies between supervised learning (pairwise labelled inputs and outputs) and unsupervised learning (completely unlabelled data).

Reinforcement learning: Reinforcement learning is related to agents attempting to maximize the overall reward in interaction with uncertain and complex environments. It is often referred to as
theoretical dynamic programming in the area of control and operational science. Like regular supervised learning, reinforcement learning does not have the right pairs of input / output and sub-optimal behaviour. In general, two approaches are commonly used to address problems of Reinforcement learning. Figure 1 depicts the survey of rural India overtaken urban India in heart-disease related deaths.

![Figure 1: Age-standardized mortality rates for coronary heart disease](image)

II. LITERATURE REVIEW

Mahmoud, et al. [1] introduces a strategy for forecasting cardiac failure utilizing global risk factors. The approach includes the two most common methods for data processing, neural networks and genetic algorithms. The hybrid method applied utilizes the regional enhancement benefit of the genetic algorithm to assign neural network weights. Training is quicker, more reliable and more precise relative to back propagation. Matlab was used to forecasts the likelihood of cardiac attack with a precision of 89%. In [2] S. K. Gaikwad proposed a theoretical method focused on the coactive neuro-fuzzy inference framework (CANFIS) for the prediction of cardiac disease was introduced. The proposed CANFIS model incorporates adaptive neural network capability with a fuzzy logic conceptual methodology, which is then combined with the genetic algorithm to diagnose the existence of the disease. Efficiency of the CANFIS model has been tested in terms of training output and detection precision, and the findings have shown that the current CANFIS model has significant potential for forecasting cardiac disease.

Deeraj, et al. [3] demonstrated a system to conduct the collection of features and the specification of parameters in an evolutionary fashion. Wrapper approach was used to choose a subset function since the values are very precise in nature. Efficiency of the ANFIS classifier was assessed in terms of training output and accuracy of the classification. Their system aimed to refine parameters and subset functions at the same time without reducing the precision of the ANFIS classification. To check the feasibility of the suggested method, the dataset for ovarian cancer was checked. The current clinical decision support method in [4] for risk management for cardiac patients consists of two phases:

1. An automatic solution for producing weighted fuzzy laws, and
2. The implementation of a fuzzy rule-based decision support framework.

In the first step, the mining process, the collection of attributes and the weighting approach were used to acquire the weighted fuzzy laws. The fuzzy structure is then designed in conjunction with the weighted fuzzy rules and the chosen attributes. Eventually, the testing was carried out on the proposed framework utilizing datasets from the UCI library and the performance of the method is compared to the neural network-based system utilizing precision, responsiveness and specificity.
S. Divyashree et al.[5] suggests a program that identifies and defines skin sores as kindly or threateningly based on photographs obtained from general cameras. Acquired photos are split, the highlights are distinguished by implementing the ABCD rule and the Neural Network is prepared to order casualties to a high degree of accuracy. Optimized Neural Network achieved an overall accuracy of 76.9 per cent for a dataset of 463 pictures, separated into six groups. The general accuracy performance and implementation of the process can be improved by training the neural network for a much broader and diverse data set with strong intra-class changeability. The misclassification which would have a positive impact on the accuracy score. The alternative, as opposed to expanding the data collection, is to increase the sum of highlights omitted from the images.

Fahad Kamal et al.[6] provided a skin tumor classification structure and explores the interaction of the skin growth image to the neural network through various forms of pre-treatment. The captured image is inserted into the photo and the pre-processing picture is used for clamor evacuation. Pictures include a segment using thresholding. There is likely to be a specific aspect in the skin growth region where such components are distinct using the Highlight Extraction Process. Multilevel 2-D disintegration of the wavelet is used to illustrate the extraction method. These highlights are supplied to the data centers of the neural network. The back-spread neural system and the out-spread vital neural system are used for classification purposes, which identify the given pictures as harmful or non-carcinogenic.

Smartphones presume actual e-well-being function such that m-well-being plays a crucial position in the human resources industry. Picture management devices are important in the medical services industry to identify anomalies in the human body. Skin development (melanoma) is one of the most harmful illnesses, but when diagnosed early, it may very well be recovered. Statistics claim that more than millions of people are suffering because of Skin Tumor itself. Sanjana M et al.[7] explores how skin disease can be differentiated from the early usage of mobile phone applications through disease dissecting features, asymmetry, boundary, color range, diameter, and extension (ABCDE). Such structures are dissected utilizing distinctive picture preparation techniques such as Gray Scale Shift, Segmentation, Follow-up Process and Histogram Analysis.

In [8], distinctive advanced pictures have been split down, given the techniques of unsupervised separation. Highlight extraction techniques are then linked to these portioned pictures. Subsequently, a specific trade was explored, based on the findings obtained. In the automatic diagnosis of skin lesions, the extraction function is based on derma-scale ABCD law. ABCD describes the asymmetry, boundary composition, colour variance and form of the derma scope referred to as the lesion diameter which determines the criteria for a dermatologist's diagnosis. At a period where a skin sore is suspected of becoming a melanoma, any one of the four checks will be performed. On the unlikely possibility that the suspected skin sore went through all three of these, it might or might not give the incorrect findings as to whether it is melanoma. Therefore, all four steps will be weighed in order to determine whether or not a skin sore is a melanoma. All things considering, the most suitable solution to getting down the risk of melanoma is to confine the exposure to strong daytime and some wellspring with Clear light. Act with the relevant figures, for example, checking that the skin is covered, using a scarf, use a sunscreen, staying in the shade (and so on). In addition, vigilance will stay on the skin and take months to months of skin self-examination to minimize the possibility of forming any skin tumor that presents a danger to human life.

In [9], Patel, proposed a system that falls under the data mining umbrella. To predict the economic activities, the program conducts data mining on trends and correlation. K-Nearest
neighbour is used in this method to estimate values that will maintain a strategic buffer from financial instability and bankruptcy. For economic estimation, classification methodology has been studied in the current analysis k-Nearest neighbour. Lately, the quantity of bankrupt companies has risen following the situation of global financial emergency. Since the financial distress of companies is the key step of bankruptcy, the use of financial proportions to predict financial distress has drawn the scholastics and economic and financial institutions into consideration.

A factual machine learning algorithm Naive Bayes (NB) is used for ordering Arabic Web documents was proposed by Sangeetha, et al.[10]. The method uses K-Nearest Neighbours to estimate values that will preserve a strategic buffer from bankruptcy and financial distress. Characterization technique for economic estimation has been investigated in the current study k-Nearest Neighbours. Lately, the quantity of bankrupt companies has risen following the worldwide financial emergency situation. Since the financial distress of companies is the key step of bankruptcy, the use of financial proportions to predict financial distress has drawn the scholastics and financial and economic institutions into consideration.

In [11] Sarvesh Chowkekar proposed a framework which attempts to explain what happens when the dimensionality increases. As dimensionality constructs the distinction between the nearest and the farthest point is distinctly insignificant and the execution is affected in this manner. This may cause false forecasting. Additionally, as much as may be predicted increment of calculation should be excluded. The k-Nearest Neighbours algorithm (or k-NN for short) is a non-parametric technique used for order and relapse, as an example of recognition. Of all machine learning algorithms, the K-NN algorithm is one of the simplest. Also for order and recurrence, relegating weight to neighbouring obligations may be beneficial, so that the near neighbours contribute more to the usual than the more distant ones.

Chen et al.[12] suggested a program that implements the nearest neighbour generation and correct algorithm configuration very quickly. In this method, they have built up a framework that allows a suitable algorithm in view of the bolstered data that rather creates the nearest neighbour as fast as possible. This algorithm is selected according to data dimension. The most boring section for certain Computer vision problems consists of the nearest neighbour coordinating in high-dimensional spaces. There are no known appropriate algorithms to fix these high-dimensional issues which are faster than straight forward pursuit. Rough algorithms are known to accurately furnish expansive speedups with only minor misfortune, but numerous such algorithms were distributed with only negligible direction when selecting an algorithm and its parameters for any given problem.

An algorithm-based image segmentation system, Support Vector Machine (SVM), and Snake Active Counter was proposed by Al-Ayyoub [13]. SVM is used to help identify the best parameters for snake algorithms. To effectively execute the snake algorithm, the initial curve and the snake parameters have to be chosen appropriately. Hence, they applied Support Vector Machine (SVM) to pick the correct initial curve and parameters. It is expected that the initial curve is of the following shapes: arc, eclipse and rectangle. These shapes were chosen to reduce the difficulty of implementing the SVM model and to preserve the SVM implementation without any degradation. Based on the results of the SVM, the attributes can be used in the image. The photos used as the test set were chosen to be used for the reference circle and to assess the accuracy of the edges. The experimental results show that the Snake algorithm is the basis for determining the edge as compared to the edge of the experts.

In [14], Ashish Kailash, et al. proposed an automatic system for the diagnosis of skin cancer from images that are clear, taken from the affected region. Melanoma, the final stage of skin cancer,
will be treated as early as possible for patients to survive. Since skin cancer occurs externally in the body, computer-aided diagnostics using digital image recognition can enable skin doctors to better diagnose skin cancer. The procedure used to diagnose melanoma are: Using the Grab Cut algorithm, the first section of the image tends to be melanoma, the next characteristics, such as structure, colour and form, are retrieved using image processing techniques. The derived characteristics are marked as cancerous and non-cancerous. Cancer called "malignant" or non-cancer mole called 'benign' by adding Gaussian radial base kernel (SVM-RBF) help vector machine. The authors performed experiments with almost 200 photos (100 of melanoma and 100 of benign) and observed from the experimental findings that only six characteristics were adequate to detect melanoma. The drawback of this approach is that the archive does not consist of protected photos to play with dark-skinned photos. They do intend to implement the Ensemble Learning Approach in the future.

Nithya, N et al.[15] recommends the use of image segmentation using an automated approach for identifying lesions using a pixel-wise pixel-based deep-learning algorithm. The analysis was performed on two network architectures by evaluating public data by using the ISIC database to train the network and the PH2 database to prove that the system used is not relevant to the data collection. The experimental findings indicate that the proposed solution is very reliable and performs segmentation in the presence of hair and air / oil bubbles. The additional contribution of this paper is the data annotation that produces several test images and the semi-automatic implementation of the GUI. Components of a compact, real-time, non-invasive skin lesion monitoring device to aid in the prevention and early diagnosis of melanoma was implemented in [16]. The first aspect is a real-time warning to help consumers deter sunlight from melting their skin; a new algorithm is added to measure the time for skin to burn. The second section consists of a robotic imaging test, including mapping, hair identification and prohibition, accident separation, highlight extraction and characterization. A structure was generated in the application for the PDA. Test results show that the proposed system is efficient, achieving high group in accuracy. Table 1 summarizes methodologies and various algorithms suggested by various authors in the field of machine learning.
III. CONCLUSION

Machine learning is a potential tool, showing its essential role in health care. Hence, this new expertise has the potential applications in healthcare sector. In our in-depth review of the literature study, we admitted that the forecast made earlier did not use a broad dataset. A broad dataset allows improved accuracy, although the feedback mechanism is not incomplete. If we foresee, we would provide some guidance to the patient about how to manage or avoid diabetes in the case of mild symptoms of diabetes. The advise will be so that it would benefit the patient if it is implemented. Therefore, we can create a program to predict diabetic patients with the aid of the Knowledge Base, where we have a dataset of about 2000 patients with diabetes, and to make recommendations on the basis of the proximity of diabetes patients. Prediction will be rendered with the aid of two Naïve Bayes and K-Nearest Neighbor algorithms and we will also evaluate which algorithm provides better accuracy on the basis of their success factors. The program that will be built will be used in the HealthCare Sector for the Therapeutic Monitoring of Diabetes Patients.
REFERENCES

I. Al-Ayyoub, Mahmoud, et al. "Machine learning approach for brain tumor detection." Proceedings of the 3rd International Conference on Information and Communication Systems, 2012.

II. Kinge, Durga, and S. K. Gaikwad. "Survey on data mining techniques for disease prediction." Int. Res. J. Eng. Technol 5.1 (2018): 630-636.

III. Shetty, Deeraj, et al. "Diabetes disease prediction using data mining." 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2017.

IV. Dekamin, Azam, and Ahmad Sheibatolhamdi. "A data mining approach for coronary artery disease prediction in Iran." Journal of Advanced Medical Sciences and Applied Technologies 3.1 (2017): 29-38.

V. S. Divyashree, H.R. Divakar Dept. of Master of computer application, PES College of Engineering Mandya, Karnataka, India Prediction of Human Health using Decision Tree Technique.

VI. Alsheref, Fahad Kamal, and Wael Hassan Gomaa. "Blood Diseases Detection using Classical Machine Learning Algorithms." Blood 10.7 (2019).

VII. Sanjana M, Dr. V. Hanuman Kumar Department of Computer Science New Horizon College of Engineering Bangalore-560103, Karnataka, India Skin Cancer Detection Using Machine Learning Algorithm.

VIII. R.Chithrakkannan, P.Kavitha, T.Mangayarkarasi, R.Karthikeyan Breast Cancer Detection using Machine Learning.

IX. Patel, Ajad, et al. "Heart disease prediction using data mining." International Research Journal of Engineering and Technology 4.1 (2017): 1705-1707.

X. Sangeetha, G., and J. Jospin Jeya. "PREDICTION OF CARDIOVASCULAR DISEASE USING SENSORS AND TECHNIQUE OF DATA MINING." (2018).

XI. Sarvesh Chowkekar, Rohan Ringe, Viral Gohil, Naina Kaushik B.E. Computer Engg. , RGIT, Mumbai, India Heart Disease Prediction using Data Mining.

XII. Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." Ieee Access 5 (2017): 8869-8879.

XIII. Al-Ayyoub, Mahmoud, et al. "Machine learning approach for brain tumor detection." Proceedings of the 3rd International Conference on Information and Communication Systems, 2012.

XIV. Pal, Ashish Kailash, et al. "Generic Disease Prediction using Symptoms with Supervised Machine Learning." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT) 5.2 (2019): 1082-1086.

XV. Nithya, N., K. Duraiswamy, and P. Gomathy. "A survey on clustering techniques in medical diagnosis." International Journal of Computer Science Trends and Technology (IJCST) 1.2 (2013): 17-23.

XVI. Gagandeep Kaur, Anshu Sharma, Anurag Sharma Department of Computer Science & Engineering, CTITR, Jalandhar, Punjab India Heart Disease Prediction using KNN classification approach.