Manisha ma'am paper

by Manisha Tiwari

Submission date: 14-Oct-2021 07:55PM (UTC+0530)

Submission ID: 1673760864

File name: Proposal_Paper.docx (53.2K)

Word count: 3016

Character count: 16653

Disease Prediction System using Machine Learning

Vaibhav Raheja
Computer Engineering
MPSTME, NMIMS
Mumbai, Maharashtra
vaibhavraheja32@gmail.com

Viraj Shah
Computer Engineering
MPSTME, NMIMS
Mumbai, Maharashtra
virajshah.vs30@gmail.com

Mayank Shetty
Computer Engineering
MPSTME, NMIMS
Mumbai, Maharashtra
msmayankshetty99@gmail.com

Purav Patel
Computer Engineering
MPSTME, NMIMS
Mumbai, Maharashtra
ppurav79@gmail.com

Prof. Manisha Tiwari
Computer Engineering
MPSTME, NMIMS
Mumbai, Maharashtra
manisha.tiwari29@gmail.com

Abstract— Chronic diseases are defined broadly as conditions that require ongoing medical attention or limit activities of daily living or both. Chronic diseases such as heart disease, diabetes, kidney disease and pneumonia are the leading causes of death and disability in the world. Early prediction of these chronic diseases would help in saving multiple lives. It is a critical challenge to detect these diseases by regular clinical data analysis. Machine learning (ML) can bring an effective solution for decision making and accurate predictions. The medical industry is showing enormous development in using machine learning techniques. Chronic Kidney Disease will be predicted using Logistic Regression and Random Forest, Diabetes will be predicted using K-Nearest Neighbour (KNN) and Logistic Regression, Heart Disease will be predicted using Decision Tree and Random Forest Regression and finally for Pneumonia we will be using Convolutional Neural Network (CNN) and Support Vector Machine (SVM).

Key Terms— Chronic Diseases, Logistic Regression, Decision Tree, ML, CNN, SVM, KNN

I. INTRODUCTION

Machine Learning is a branch of Artificial Intelligence (AI), where the main objective is to give the computer the ability to learn from a provided set of data. The structure of the data is understood, after which the data is fit into models. These models can be successfully utilized by people for any given application where machine learning is required.

Despite being a field within computer science, it radically differs from the traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers for calculations or problem-solving, whereas Machine Learning algorithms allow computers to train on data inputs and use statistical analysis to output values that fall within a specific range. This approach in Machine Learning facilitates computers in building models from sample data, to automate decision-making processes based on data inputs.

Our project will be harnessing the potential of machine learning, in which a model will be trained in identifying various diseases included in our scope, where the output will be in Boolean values.

II. LITERATURE REVIEW

Machine learning algorithms have not only proven to be an essential tool in the field of mathematics and engineering, but also in the field of medicine as well.

The authors in paper [1] survey various machine learning algorithms and evaluate their accuracy on the Pima Indian Diabetes Database. Among the various models that are surveyed in the paper, Logical Regression and Support Vector Machine provided an accuracy of 78% in both K-fold and test-train split. All models give an average accuracy of 70%. Among all the proposed models, the neural network with two hidden layers is considered the most efficient and promising for analysing diabetes with an accuracy rate of approximately 86%. In paper [3], the authors have used a larger dataset based on the same parameters as the PIMA Indian Diabetes dataset which boosted the accuracy of the Logistic Regression classifier to 96%, alongside the pre-processing of the dataset.

The authors in paper [5] design an interface to get the user's input parameters and predict whether the user is healthy or has a probability of getting a heart disease. Decision Tree and Random Forest are used to make the prediction. With Decision tree the authors achieved an accuracy of 79% and with Random Forest they the accuracy achieved was 81%. Along with these two algorithms the authors experimented with a Hybrid model which was a combination of Decision Tree and Random Forest. The Hybrid model was the most efficient and had the highest accuracy between the three algorithms with 88.7%.

In paper [7], the authors establish different machine learning algorithms to diagnose chronic kidney disease using the UCI Chronic Kidney Disease dataset using their proposed model. The performance of several high-accuracy algorithms was analysed, after which only two were shortlisted for an integrated model, namely Logistic Regression and Random Forest. The accuracy of both these algorithms were 98.95% and 99.75% respectively, while the accuracy of the integrated model was 99.83%, which was higher than the individual accuracies of the algorithms used in the integrated model.

The authors of paper [11] utilised a pneumonia dataset from Kaggle, which consisted of chest X-ray images of several patients. The pre-processing of the dataset consisted of resizing the images and creating a function which stores the pre-processed images. The images of the chest X-ray were converted to an array of numbers, after which the authors used Convolutional Neural Network (CNN) to train the model to detect pneumonia, which ultimately provided an accuracy of 98.06%.

III. PROPOSED SYSTEM

Figure 1 shows the flowchart diagram of the proposed chronic disease detection system.

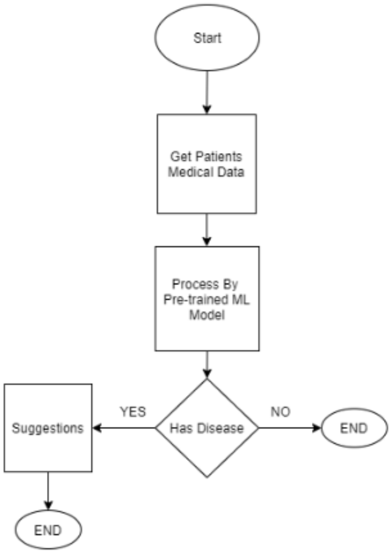

Fig. 1. Flowchart of the proposed system

A. Get Patients Medical Data

This state deals with the collection of input parameters for the pre-trained model. A specific set of biological parameters are collected and fed to the pre-trained machine learning model, depending on the selected chronic disease.

B. Process By Pre-Trained Machine Learning Model

This stage contains a pre-trained model for each chronic disease considered in our project. Appropriate features are selected for each chronic disease, which helps in further classification of a patient having or not having the specified chronic disease.

C. Suggestions

This stage is the output of the pre-trained model for the disease. In the event of a positive result, appropriate suggestions are provided to the medical expert, as there is a high probability of the patient having a chronic disease. If the result is negative, the patient has a low probability of the patient having a chronic disease.

IV. PROPOSED ALGORITHM

A. Architecture Diagram

Figure 2 shows the architecture diagram of the proposed chronic disease detection system.

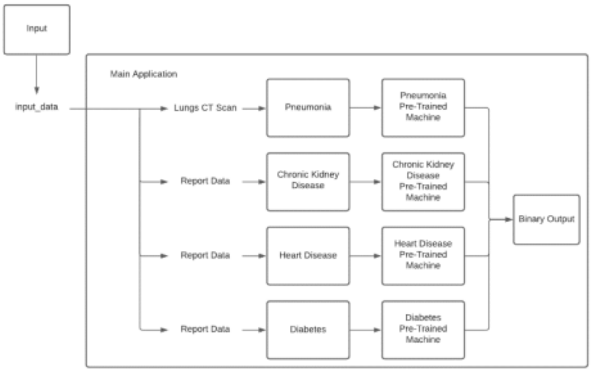

Fig. 2. Architecture diagram of the proposed system

The user will initially select the chronic disease model they want to use and upon selecting one of the specified diseases, the data is input into the chronic disease detection systems' pre-trained machine learning models. The models will then give a binary output, stating whether the patient has the selected disease or not.

The pre-trained machines will be trained using a pre-processed dataset. The dataset will go through multiple processing steps, in which the data will also be checked for any noisy or missing values before being fed to the machine learning models.

K-Nearest Neighbour (KNN) will be used for imputing missing values, outlier detection methods will be used to estimate any noise in the data, and rapid miner will be used to remove noise in the dataset. The dataset will also be checked for discrepancies using data transformation, discretization, and binning techniques.

B. *Machine learning algorithm for Chronic Kidney Disease Detection*

Logistic Regression: Logistic Regression is a statistical machine learning technique used for binary classification of categorical dependent variables, by using a set containing independent variables. Assuming the probability of a particular class, Logistic Regression creates a regression model that distinguishes between several samples using the sigmoid function, where one or more variables are used to determine the expected outcome in probabilistic values which lie between 0 and 1. Logistic Regression has previously been implemented to detect chronic kidney disease in patients, and from referenced papers it was found to have a detection accuracy of 98.95% [7], 99.24% [8], and 100% [9] respectively on the UCI dataset, making it a high-accuracy machine learning model for implementation of the expert system.

Random Forest: Random Forest is a supervised learning classification technique and is often considered as an ensemble machine learning method, as it is used for classification, regression, and probability. Random Forest can find missing values from many datasets, and it can provide a more accurate value by creating a forest of decision trees during the learning phase, where the number of trees indicate the robustness of the forest as well as the accuracy of the algorithm. For training characteristics and then random sub characteristics for sampling nodes, random sampling is done. The algorithm aggregates multiple decisions to make a single decision by consolidating multiple forests on different subsets of a dataset and averaging the results to enhance the performance of the dataset's detection accuracy. It can be combined with multiple classifiers to solve a complex problem and improve the overall accuracy of the machine.

Step 1: Randomly select 'p' features from the total 'q' features, where p << q.

Step 2: Calculate node 'd' using the method best split point from the selected 'p' features.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat the above steps till a number '1' is reached.

Step 5: A forest of 'n' number of trees is built by applying the above steps 'n' number of times.

Random Forest has been chosen as it comparatively takes less time to train and has also been used to detect chronic kidney disease in patients from our referenced papers, where it was found to have a detection accuracy of 94.16% [8], 99.75% [7], and 100% [9] respectively on the UCI dataset, making it a suitable high-accuracy machine learning model for the implementation of the expert system. It can be combined with a Logistic Regression classifier to form an Integrated Model, which offers an average detection accuracy of 99.83%, which is better than the average detection accuracies of both the algorithms combined [7].

C. *Machine learning algorithm for Diabetes Detection*

K-Nearest Neighbour: KNN algorithm stores all the available data and classifies a new data point based on similarity. This means that when any new data appears, it can be easily classified into a well-defined category using the KNN algorithm. KNN received the highest accuracy of 79% [1] in detection of diabetes on PIMA dataset.

Step 1: Importing the Dataset

Step 2: Choose the value of K that is the nearest data point.

Step 3: For each of the K in test data do:

- Calculate the distance between each K and each row of the training data.
- Sort in ascending order based on distance value.
- Choose top k rows from the sorted array.
- Assign a class to the test point based on the most frequent class of the row.

Hence, it can be used for a high-accuracy machine learning model to detect diabetes in patients, as well as a reference for further improving on the accuracy of the model.

Logistic Regression: Logistic regression, despite its name, is a classification model rather than a regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It predicts the output of a categorically dependent variable, where the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression receives an accuracy of 78% [1] in detection of diabetes on PIMA dataset. Hence, it can be used for a high-accuracy machine learning model to detect diabetes in patients. Pre-processing the data using K-means and clustering can also increase the accuracy to 96% [3], making it a suitable model for implementation of the expert system.

D. *Machine learning algorithm for Heart Disease Detection*

Decision Tree: It's a tree-like classification model, which builds a structure consisting of branches and nodes on the bases of evidence collected for each attribute during the model learning phase. The decision tree's branches and nodes connect according to the number of entities described in the dataset. The forwarding process uses the number of values dedicated for each attribute. Furthermore, following the rules described on each branch and node, it reached the decision for each transaction. Finally, according to the decision node the class label will be assigned to the record. This procedure is iterative and repeats till each transaction has a class category. Therefore, this algorithm converts the attributes into branches and nodes, and selects one of the attributes as a decision node.

Step 1: Importing the libraries

Step 2: Importing the dataset

Step 3: Splitting the dataset into the Training set and Test set

Step 4: Training the Decision Tree Regression model on the training set

Step 5: Predicting the Results

Step 6: Comparing the Real Values with Predicted Values

Step 7: Visualising the Decision Tree Regression Results

We are going to use the decision tree for Heart disease detection, as the outputs the referenced papers [6] had good accuracy, while being easy to read and interpret without requiring any statistical knowledge. It takes less effort to prepare the data, and once the variables have been created, the data cleaning process is minimal. Decision Tree algorithm achieved an accuracy of 79%[5].

Random Forest Regression: Random Forest Regression aggregates multiple decisions to make a single decision. For training the characteristics and the random sub characteristics sampling nodes, random sampling is done. It combines multiple classifiers to solve a complex problem and improve the machine's accuracy. During the learning phase, this model first generates multiple random trees called a forest. It is a classifier that consolidates multiple forests on different subsets of a dataset and averages the results to enhance the performance of the dataset's detection accuracy.

Step 1: Importing the libraries

Step 2: Importing the dataset

Step 3: Splitting the dataset into the Training set and Test set

Step 4: Training the Random Forest Regression model on the training set

Step 5: Predicting the Results

Step 6: Comparing the Real Values with Predicted Values

Step 7: Visualising the Random Forest Results

Random Forest Regression was selected from the referenced papers [6], as it comparatively takes less time to train while providing high accuracy, which increases the efficiency of the model. Random Forest Regression achieved an accuracy of 81% [5].

E. Machine learning algorithm for Pneumonia

Convolutional neural network: The system learns to perform feature extraction in a convolutional neural network. The core concept of CNN is to use convolution of image and filters to generate invariant features which are passed into the next layer. The characteristics of the following layer are combined with different filters to produce more invariant and abstract features. This is done till the final feature / output (say face of X) is achieved. Convolutional neural network consists of several building elements such as convolution layers, pooling layers, and fully connected layers, and is meant to learn the spatial hierarchies of features by means of a background algorithm

.

Support Vector Machine (SVM): The objective of the SVM algorithm is to find the optimal line or decision boundary for categorising n-dimensional space to categorize new data points in the correct category. The best decision boundary is referred to as a hyperplane.

Step 1: Import Python libraries

Step 2: Display image of each bee type

Step 3: Image manipulation with rgb2gray

Step 4: Histogram of oriented gradients

Step 5: Create image features and flatten into a single row

Step 6: Loop over images to pre-process

Step 7: Scale feature matrix + PCA

Step 8: Split into train and test sets

Step 9: Train model

Step 10: Score model

Step 11: ROC curve + AUC

V. CONCLUSION

Expert systems in medicine are defined as systems with the ability to capture and store expert knowledge, facts, and reasoning techniques to assist doctors in diagnosing a patient's condition. These systems attempt to mimic a doctor's expertise by applying several computational methods to help in decision support and problem solving, by coming up with reasoned conclusions for a patient's illness or condition. Our project will incorporate the core elements of an expert system by supporting medical experts with their claims, precursing their diagnosis of a chronic disease in their patients using trained machine learning models giving high accuracy of disease detection and prediction.

VI. REFERENCES

[1] J. J. Khanam, S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction", *ICT Express (Feb. 2021)*, 2021, doi: 10.1016/J.ICTE.2021.02.004.

[2] D. Shetty, K. Rit, S. Shaikh and N. Patil, "Diabetes disease prediction using data mining," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1-5, doi: 10.1109/ICIIECS.2017.8276012.

[3] A. Mujumdar, V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms", *Procedia Computer Science, Volume 165*, 2019, pp. 292-299, doi: 10.1016/J.PROCS.2020.01.047.

[4] S. K. Dey, A. Hossain and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 2018, pp. 1-5, doi: 10.1109/ICCITECHN.2018.8631968.

[5] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[6] F. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease", *International Journal of Advanced Computer Science and Applications (IJACSA), 10(6)*, 2019, doi: 10.14569/IJACSA.2019.0100637.

[7] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," in *IEEE Access*, vol. 8, pp. 20991-21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[8] R. Gupta, N. Koli, N. Mahor and N. Tejashri, "Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease," *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154147.

[9] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298026.

[10] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1-7, doi: 10.1109/ICECCT.2019.8869364.

[11] A. Sharma, M. Negi, A. Goyal, R. Jain and P. Nagrath, "Detection of Pneumonia using ML & DL in Python.", *IOP Conference Series: Materials Science and Engineering* 1022, 2021, doi: 10.1088/1757-899X/1022/1/012

Manisha ma'am paper

ORIGINALITY REPORT

43%	29%	33%	26%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	thesai.org Internet Source	5%
2	M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj. "Heart Disease Prediction using Hybrid machine Learning Model", 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021 Publication	4%
3	www.upgrad.com Internet Source	4%
4	www.c-sharpcorner.com Internet Source	3%
5	www.javatpoint.com Internet Source	2%
6	www.datacamp.com Internet Source	2%
7	ijircce.com Internet Source	1%

8	"Soft Computing and Signal Processing", Springer Science and Business Media LLC, 2020 Publication	1%
9	Abdulhamit Subasi. "Machine learning techniques", Elsevier BV, 2020 Publication	1%
10	Submitted to Aspen University Student Paper	1%
11	Mercedes Rigla, Gema García-Sáez, Belén Pons, Maria Elena Hernando. "Artificial Intelligence Methodologies and Their Application to Diabetes", Journal of Diabetes Science and Technology, 2017 Publication	1%
12	Rahul Gupta, Nidhi Koli, Niharika Mahor, N Tejashri. "Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease", 2020 International Conference for Emerging Technology (INCET), 2020 Publication	1%
13	Amit Birajdar, Harsh Agarwal, Manan Bolia, Vedang Gupte. "Image Compression Using Run Length Encoding and Lempel Ziev Welch Method", 2019 Global Conference for Advancement in Technology (GCAT), 2019 Publication	1%

14	Submitted to International University of Malaya-Wales Student Paper	1%
15	Submitted to University of Hertfordshire Student Paper	1%
16	www.ijariit.com Internet Source	1%
17	Submitted to Coventry University Student Paper	1%
18	Jobeda Jamal Khanam, Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction", ICT Express, 2021 Publication	1%
19	Submitted to University of Northampton Student Paper	1%
20	info.cloudquant.com Internet Source	1%
21	dokumen.pub Internet Source	1%
22	www.ijert.org Internet Source	1%
23	Submitted to Institute of Research & Postgraduate Studies, Universiti Kuala Lumpur Student Paper	1%

24	Submitted to Cornell University Student Paper	1%
25	"Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2019 Publication	<1%
26	Submitted to University of Reading Student Paper	<1%
27	www.ijrte.org Internet Source	<1%
28	medium.com Internet Source	<1%
29	Penghe Liu, Xiaoqing Wang, Xiaoping Sun, Xi Shen, Xu Chen, Yuzhong Sun, Yanjun Pan. "Chapter 8 HKDP: A Hybrid Knowledge Graph Based Pediatric Disease Prediction System", Springer Science and Business Media LLC, 2017 Publication	<1%
30	A Sharma, M Negi, A Goyal, R Jain, P Nagrath. "Detection of Pneumonia using ML & DL in Python", IOP Conference Series: Materials Science and Engineering, 2021 Publication	<1%
31	www.ijeat.org Internet Source	<1%

32	Hoda Ramin Hossein, S. S. Shaikh. "SPHPMS: Smart personnel m-healthcare patient monitoring system", 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016 Publication	<1 %
33	J. Jospin Jeya, E. Kannan. "Multi Key Word Search and Trusted Auditing System to Verify the Integrity of Outsourced Data in Cloud Computing", International Review on Computers and Software (IRECOS), 2014 Publication	<1 %
34	Prasannavenkatesan Theerthagiri, Usha Ruby A, Vidya J. "Diagnosis and Classification of the Diabetes Using Machine Learning Algorithms", Research Square Platform LLC, 2021 Publication	<1 %
35	Submitted to City University Student Paper	<1 %
36	christuniversity.in Internet Source	<1 %
37	etd.astu.edu.et Internet Source	<1 %
38	Submitted to iGroup Student Paper	<1 %
39	towardsdatascience.com Internet Source	

<1%

40 Rajat Yadav, Anurag Mishra. "Identifying and sensing emotional quotient weightage in the outcome using Speech Dialogue", IOP Conference Series: Materials Science and Engineering, 2021
Publication
<1%

41 "Proceedings of Second Doctoral Symposium on Computational Intelligence", Springer Science and Business Media LLC, 2022
Publication
<1%

42 Bart Selman. "ExOpaque: A Framework to Explain Opaque Machine Learning Models Using Inductive Logic Programming", 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007), 10/2007
Publication
<1%

43 Meenakshi Garg, Gaurav Dhiman. "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants", Neural Computing and Applications, 2020
Publication
<1%

44 Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat et al. "Prediction of Chronic Kidney Disease -A
<1%

Machine Learning perspective", IEEE Access, 2021

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		