Smart System to Examine Heart Disease and Lung Cancer Using Machine Learning Methods

[1]ATHARV ARYA, [2]AYUSH SINGH, [3]AMAN LAKHMANI

[1,2,3]Electronics and Communication Engineering Department VIT India
Email: [1]rishu4321@gmail.com, [2]ayushsingh29874@gmail.com, [3]lakhmaniaman9@gmail.com

Abstract: In clinical data analysis, predicting cardiovascular disease and lung cancer is a primary challenge, nowadays, patients face a situation of unanticipated demise due to specific heart and lung problems/diseases that are either left undiagnosed or too late to get detected. Recent advances in healthcare technologies have introduced new means of diagnosis of these diseases, one being through Machine Learning (ML). The proposed paper uses machine learning algorithms to predict heart disease and lung cancer based on various parameters along with better visualization. We have used a benchmark dataset of UCI Heart Disease Prediction which has 14 different parameters and the lung cancer dataset consists of 16 different parameters to predict the disease. Both diseases are predicted using different algorithms and the accuracies are compared, the best one is taken out. Since not every person is trained enough in using machine learning methods to input the parameters and get the results we incorporate the whole method into a mobile application for better visualization and ease of access.

Keywords: Machine Learning, Cardiovascular Disease, Heart Disease

1. INTRODUCTION

Due to increasing environmental degradation, the quality of air is going down which is resulting in increasing lung and heart diseases. Lung cancer, is a leading cause of death worldwide, starting from the lungs and spreads to other organs of the body and has a low survival rate of just 15%. It is estimated that at least 1.2 million people were infected with the disease, equivalent to 12.3% of the total number of cancers diagnosed annually, with a mortality rate of 1.1 million people per year. In addition, heart disorder is the principal motive for a large number of deaths inside the world over the last few years and has emerged as the most existence-threatening sickness, now not only in India but within the entire world. So, there may be a need for dependable, accurate, and feasible gadgets to diagnose such illnesses in time for correct remedy. gadget studying algorithms and strategies were applied to numerous scientific datasets to automate the evaluation of big and complicated records. This is due to different risk factors such as dietary habits, physical inactivity, alcohol consumption, among others.

According to World Health Organization statistics, 4.9 million people die of lung cancer from snuff intake, 2.6 million are overweight, 4.4 million have elevated cholesterol, and 7.1 million have high blood pressure. Chronic disease deaths are said to increase by 17% over the next 10 years, which translates to around 64 million people. Chronic diseases vary a great deal in their symptoms and how they evolve and are treated. If not treated early, some cancers can kill a patient. The most common chronic diseases that can be treated and monitored are diabetes, blood pressure, and cardiac arrhythmias. Patients with these illnesses often have not only a limited physical condition, but also financial, emotional, and social relationships, among others. The idea of the net of factors is the latest and is defined as the combination of all devices that hook up with the community, which can be managed from the web and in turn offer facts in actual time, to allow clean access and a user-friendly interface. Any other concept of IoT "is the overall concept of things, especially ordinary objects, which can be readable, recognizable, locatable, addressable and controllable thru the net - both through RFID, wireless LAN, wide region network, or through other means ". The data obtained from the conventional sensors are not very useful because each sensor reads different aspects and gives different data and because of the absence of a common platform, it is harder to take all the readings into account and predict it. These predictions are done by the doctors which are not possible when the patient is not at the hospital under doctors' supervision. Several parameters have been identified that cause heart disease in which some of which are also common in lung disease along with some other extra parameters. These parameters are used by experts to identify the disease but due to increasing health risks and different parameters being identified by the new technologies it has become harder for healthcare workers except for the experts.

Using the IoT interoperability and more connected devices with increasing technology we now can obtain different parameters sensed by different sensors onto a single platform without using complex operations which a normal person would get confused while using. The digital system modern hospitals use now produce an enormous amount of data that is hard to track, we now know that several parameters can be used to detect diseases for early diagnosis. This factor led to research on the processing of medical pictures Due to the lack of experts and the number of cases incorrectly diagnosed, a rapid and efficient automated detection system was required.

Here we combine two important aspects of the chest, namely lungs and heart, using different datasets for heart disease and another dataset for lung disease. This will help an individual to get full information about the chest in one place. Using

machine learning and different methods to get a good accuracy for a better diagnosis and a user-friendly interface for normal people to interact with.

2. LITERATURE SURVEY

In this literature various machine learning algorithms and deep learning-based diagnosis techniques have been proposed to diagnose diseases related to the cardiovascular systems. This research study presents some diagnostic techniques based on machine and deep learning to explain the significance of the proposed work.

Detrano et al. [1] developed a heart disease classification system by using machine learning classification techniques and the overall accuracy of the system was 77%. In this paper, the Cleveland dataset was used with the features selection method.

In another study, Guo et al. [2] introduced/developed detection of heart disease by using machine learning techniques namely Recursion enhanced random forest with an improved linear model (RFRF-ILM). This paper focuses on the detection of Heart Disease using a machine learning model, the proposed RFRF-ILM method is applied by merging the features of the random forest and linear model. The proposed algorithm in this paper saves overall cost and time for the diagnostic and is returned with an accuracy of 96.6%.

Li et al. [3] designed an efficient machine learning-based diagnosis system. This study proposed multiple Machine Learning classifiers which include LR, K-NN, ANN, SVM, and DT are used in designing the system. The proposed diagnosis system achieved good accuracy as compared to other past methods.

Hosseinzadeh et al. [4], This paper predicts several types of lung tumours based on protein attributes by machine learning algorithms. Methods such as Feature extraction and feature selection process are used to detect the two main types of cancer and take 12 different parameters
into account. The performance also increased when using weighing models instead of the original datasheet. This Machine learning consists of seven SVM models, three ANN models, and two NB models which make the predictions have different weights according to their seriousness. After running the models, the SVM dataset gave the best accuracy of 88% and showed that using feature extraction and selection process significantly increased the accuracy. Using the right weights on different types of accuracy also has a significant effect.

R. Thomas et al. [5], In this paper the author discusses about outliers in the dataset that we use and its importance in making machine learning based software . He also discusses about outlier detection algorithm like one class SVM and auto encoder that will help in more precise training of our data.

C. Gao et al. [6], the author uses Logistic regression and use it with clustering algorithm on the continuous healthcare data and shows that this approach can increase precision and model accuracy.

El et al. [7], author discussed various machine learning (ML) algorithms, it's background and how these algorithms works and its application in medical physics and radiation oncology.

The paper by Trusculescu et al. [8] shows Interstitial lung disease refers to a group of over 100 lung disorders. This paper uses CNN in deep learning to identify the type of disease for early diagnosis. high-resolution tomographic images are used in pattern recognition, and some similar types of disorders were misclassified. Conventional accuracy was around 82.1% but with some specific datasets, it increased up to 89%. The algorithm also gave the best results similar to the human capacity for some of the disorders but the main drawback was that its iterative algorithm required more resources which is not available in normal computers.

S. Daberdaku et al. [9], This paper discusses about the K-nearest neighbor algorithm and how to improve it based on the real world dataset because many data have missing values and to calculate the algorithm needs full data. The incomplete data set may give biased and wrong predictions.

W. Xing et al. [10] author discussed about KNN, it's simplicity for classification of big medical health data and also proposed an improved KNN algorithm, later comparison between proposed and traditional algorithm.

Another paper by Nisar et al. [11] discusses many Machine Learning and Deep Learning Models covering both supervised and unsupervised learning and their use and accuracy in healthcare fields.

R. Lee et al. [12], While other papers discuss about the prediction of disease, this paper discusses about a specific disease called diabetes and tells whether the person has diabetes or not based on the accurate data set. the system in the paper will discuss about two types of diabetes. The paper also works on 5 types of prediction algorithms which are-Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, Decision Tree and Random Forest algorithm.

P. Amudha et al. [13] In this paper the author discusses about the healthcare industry requires high volume of data for preprocessing and for that he suggests the use MapReduce to store data in less computation. This approach is very useful in big data analysis.

S. Tayeb et al. [14],In this paper the k-nearest neighbors algorithm was used to deal with large amount of data. As the large amount of data contains valuable information it cannot be discarded but it is harder to deal with, using this we get accurate and efficient results. This learning method was applied to the dataset provided by University of California about two diseases - chronic kidney failure and heart disease and the accuracy from k-nearest neighbors algorithm was found to be 90%.

A. Singh et al. [15], In this paper author calculated the accuracy for their proposed ML algorithms by using UCI repository dataset for training and testing, in which they used jupytr notebook for implementing the python program.

Sunita et al. [16] In this paper, the automatic detection of patterns of intermediate lung disease in high-resolution computed tomography images is achieved by constructing a network-based network detector with GoogLeNet as the backbone. GoogLeNet has been simplified by releasing the first few models and used as the backbone of the detector network. The proposed framework was developed to identify several intermediate lung disease patterns without classification of lung areas. The proposed method is able to identify five of the most common patterns of interstitial lung diseases: fibrosis, emphysema, consolidation, micronodules and low-grade visual acuity, as well as generalized.

A. D. Gunasinghe et al. [17]In this paper the author uses machine learning and deep learning to create a model using process data by combining the patient information of their corresponding chest x-rays, CNN is used with an existing pre-trained model. Finally prediction is done for breathing problems such as Asthana, chronic obstructive pulmonary disease (COPD), Tuberculosis, Pneumothrax and Lung cancer

Angelini et al. [18] This paper aims to discuss current challenges and potential for artificial intelligence (AI) in lung fungus, focusing on chronic aspergillosis of the lungs and others that support the results of psychological evidence using lung imaging.

Kwekha et al. [19] The purpose of this study was to discover the role of machine learning applications and algorithms in the investigation and various objectives related to COVID-19. In this paper Supervised learning has shown better results than other non-supervised learning algorithms with 92.9% test accuracy.

Krishnan et al. [20] This is especially true for elderly patients. In this paper they propose a new plan to escape the rate of sudden death through Patient Health Monitoring which uses sensory technology and uses the internet to communicate with loved ones in case of problems. This system uses temperature sensor and heart to track the health of patients. Both sensors are connected to Arduino-uno.

The paper forms the groundwork for our research by differentiating the Machine and Deep learning algorithms such as MLP, Auto-encoder, SVM, CNN, etc. according to which algorithm best fits as a learning model for a particular type of disease whether it may be CNS, cardiovascular system, and respiratory system, This paper also emphasizes on the use of better and clean dataset to be used in ML and DL models as they help the model to learn and improve faster.

3. PREDICTION MODEL

We have considered different datasets for lung cancer and heart disease, there are 15 attributes with a dataset size of 310 for lung cancer as shown in Table1, and 13 attributes with a dataset size of 304 for heart disease shown in Table2. The total patient records amount to 614 for both the diseases.

Different algorithms give prediction with different accuracies and so we choose the one with the maximum accuracy.

Attributes		
Age	Sex	Chest Pain Type
Resting Blood Pressure	Cholesterol	Fasting Blood Pressure
Resting Electrocardiographic Measurement	Max Heart Rate Achieved	Exercise Induced Angina
ST Depression	Slope	Thalassemia
Number of Major Vessels		

Table.1. Accuracy of various algorithms for heart disease.

Attributes		
Age	Sex	Smoking
Yellow Fingers	Anxiety	Peer Pressure
Chronic Disease	Fatigue	Allergy
Wheezing	Alcohol Consuming	Coughing
Shortness of Breath	Swallowing Difficulty	Chest Pain

Table.2. Accuracy of various algorithms for lung cancer.

The prediction model is given as:

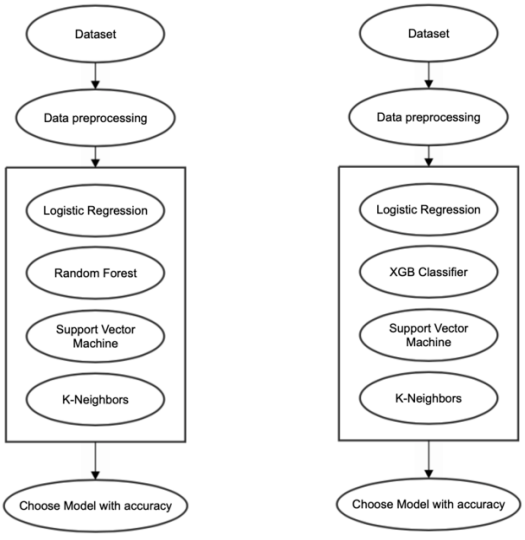

Fig.1. Prediction Model

A. *Input*

The input data will be entered by the user which he is experiencing among the attributes, this data will be processed by the algorithms.

B. *Data pre-processing*

Data pre-processing is a data mining and data analysis technique that takes in the raw data and transform it

into a suitable format for machine learning. These were the techniques used in this work:

- Data Cleaning: One of the most important step of pre-processing, Data obtained from sensors and other medical equipment are not always clean and contain irregularities, so to clean and present data in formatted way we perform data cleaning.

- Data Transformation: In this process data is transformed into a better-organized format as to improve the data quality. There are many steps involved such as normalization, filtering, feature selection

C. *Models selected*
For heart disease, Our system will predict the disease using

• Logistic regression
• K-Nearest Neighbors
• Random forest
• Support Vector Machine

For lung cancer, the system uses following algorithms

• Logistic regression
• K-Nearest Neighbors
• XGB classifier
• Support Vector Machine

The accuracy will be compared in the results and implementation section in this paper.

D. *Output*
After the training of the system using the algorithms, a model is formed. The given user input data is tested against this formed model and the probability of having a disease is given as output.
So the accuracy of a certain algorithm is very crucial which would help us to predict a disease so that the system is reliable.

4. METHODOLOGY

4.1. Support Vector Machine

A support vector machine is a supervised machine-learning method that can be used in both classification and regression problems. It's primarily used for classification problems in Machine Learning. The goal of the SVM is to find a hyperplane in an N-dimensional space where N is the number of features, that separates the data points into two classes and maximises the marginal distance for both classes, as a larger margin will produce a lower Vapnik–Chervonenkis (VC) dimension, which will reduce the possibility of generalisation error. The marginal distance is defined as the distance between the hyperplane and its nearest data points which is a member of that class. Fig.2 shows the working of SVM.

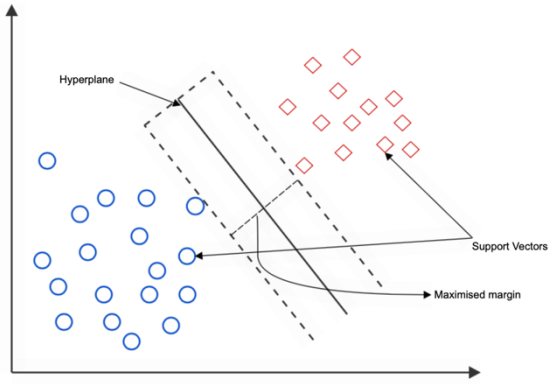

Fig.2. Illustration of how the support vector machine works, it finds a hyperplane which maximises the separation between 'circle' and 'diamond' classes.

4.2. Logistic Regression

Regression analysis is a type of modelling technique which finds the relationship between one dependent variable and one independent or multiple independent variables. When multiple variables are used out is known as multiple regression.
There are two types of regression :
linear regression and logistic regression.
Logistic regression is a type of regression analysis used for prediction for the outcome of a certain dependent variable from the set of independent variable which are also called predictor elements.
LR is mostly used for binary classification problems. Since the dependent variable is always binary so it is used for prediction and probability of failure and success. It is most suited where the dependent variable is dichotomous, which means that there can only be two outcomes.

4.3. K-Nearest Neighbors

It is a type of classifier that stores all the data and then classifies the new data on the basis of similarity between the data, shown in Fig.3 basically the data is classified on the basis of how the neighbour data is classified.
This type of supervised machine learning algorithm can also be used for regression predictive problems apart from classification.
The value of K is the number of datapoints we want initially. KNN loads the dataset and uses a feature similarity to predict the new data points. For each data point: The Euclidean distance of the data point to all the training data is calculated, the data is sorted into ascending order and top K entries are chosen.
This way the data which have similar attributes are assumed to exist in close proximity. The value of K here is important and has to be chosen wisely.

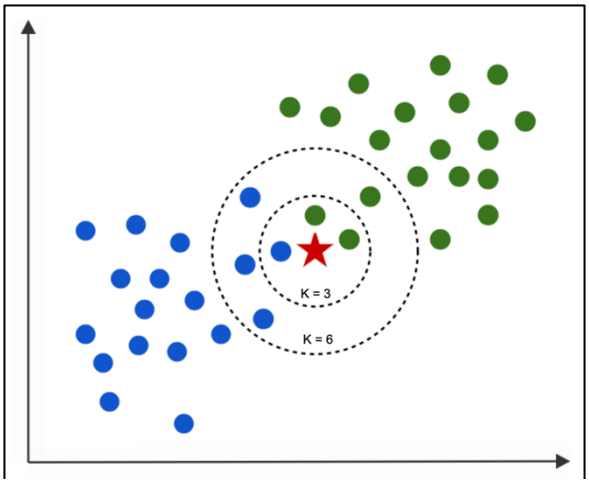

Fig.3. Illustration of K-nearest neighbor algorithm. When K = 3, sample (star) is classified as class 'green'. For K = 6, the same sample object is classified as class 'blue'.

4.4. Random Forest

Random forest is a type of ensemble classification and it the most effective approach. The RF is mostly used in prediction and probability estimation.

Random forest contains a large number of decision trees and each tree gives a vote that have an impact on the decision about of the class of the object shown in Fig.4.

Random forest is highly advantageous where the data is in large amounts as it can handle hundreds of input variables.

RF can also estimate which data is important while calculations, it also handles the missing entries in the given dataset.

It uses random selection and bagging of the features which has three important tuning parameters:

1- No. of trees

2- Minimum node size

3- No. of features employed in splitting each node for each tree.

4.5. XG Boost classifier

XGBoost also known as Extreme Gradient Boosting is a serial boosting algorithm written in C++ proposed by T. Chen & C. Guestrin in 2016. XGBoost is a decision-tree-based ML algorithm that uses CART(Classification and Regression Trees) tress instead of traditional random forest trees shown in Fig.5, and uses a gradient boosting framework, built to increase the ml model performance and computational speed, as in this algorithm trees are built in parallel in comparison to sequential trees in GBDT.

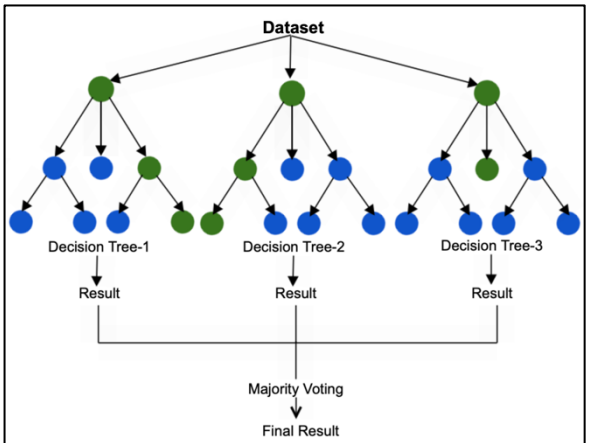

Fig.4. Structure of Random Forest which consists of three different decision trees.

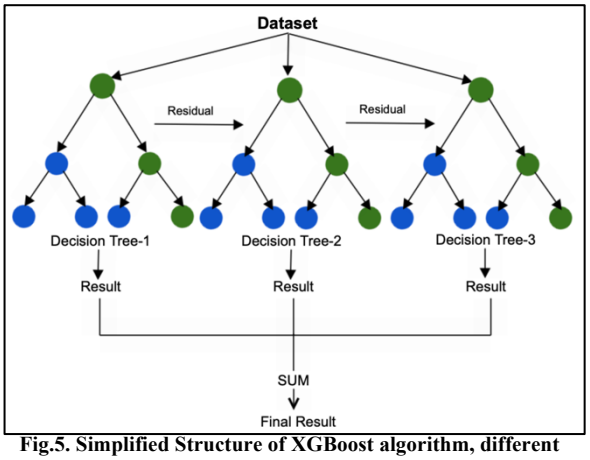

Fig.5. Simplified Structure of XGBoost algorithm, different from random forest, each decision tree minimises the residual from its previous model.

5. RESULTS AND IMPLEMENTATION

5.1. Performance of Algorithms

With the increasing toxicity in our day-to-day lives and even in the air we breathe there are several health concerns. These are also one of the leading main causes of death worldwide. And since lung cancer and heart disease are detected after the damage has been done, early detection is very necessary for it to be diagnosed. There are various small symptoms that a person gives early in the disease which when taken in a composed way we can detect the disease early. This is done using machine learning methods and different algorithms are developed by different people to take all the different parameters into account along with their weights. First, the algorithm is trained using the existing dataset, and the parameters are entered into it, then the result given by the algorithm is compared against the results given in the dataset and the accuracy is calculated.

The accuracy comparison for different algorithms for heart disease is given in table 3. It shows that the Logistic Regression and Random Forest model gives the highest accuracy.

S.No.	Algorithms	Accuracy Score

1	Logistic Regression	88.52
2	K-Nearest Neighbors	86.88
3	Random Forest	88.52
4	Support Vector Machine	85.24

Table.3. Accuracy of various algorithms for heart disease.

The accuracy comparison for different algorithms for lung cancer is given in table 4. It shows that the Logistic Regression model gives the highest accuracy.

S.No.	Algorithms	Accuracy Score
1	Logistic Regression	91.07
2	K-Nearest Neighbors	87.50
3	XGB Classifier	89.28
4	Support Vector Machine	82.14

Table.4. Accuracy of various algorithms for lung cancer.

5.2. Implementation

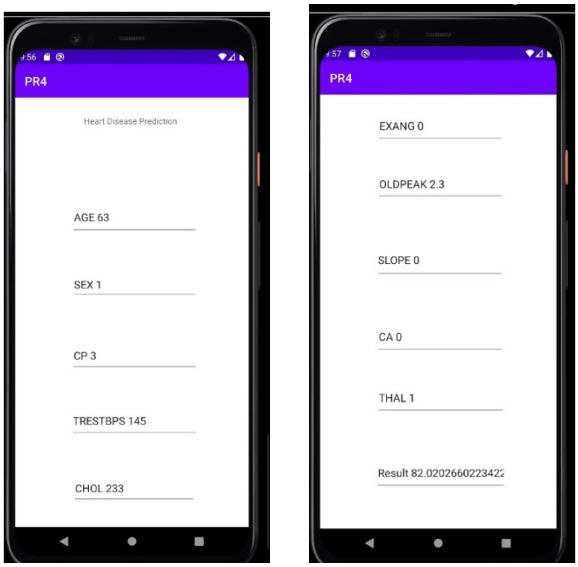

Fig.3. App Interface

Fig.2 shows the app simulation based on the proposed methodology for heart disease prediction. It takes the parameter from the user and gives the approximated probability at which heart is at risk based on the logistic regression algorithm. In this app, we used Logistic Regression other than other proposed algorithms so that we don't have to make an extra ask file for machine learning purposes. Using the weights in Logistic Regression algorithm and normalizing our data we got better time complexity for the app and app is lightweight.

CONCLUSION

In this paper, we introduced four algorithms in which comparative analysis was done and promising results were achieved. The correct prediction of heart disease can prevent life threats, and incorrect prediction prove to be fatal. In this paper different machine learning algorithms are applied to compare results and analysis of Machine Learning Heart Disease dataset. After comparing the machine learning models we introduced to predict lung cancer and heart disease based on the parameters provided we came up with the model that works best on a dataset of a certain area. Logistic regression gives a better result than other alternatives. The dataset consists of 10 attributes used for performing the analysis. Using machine learning we obtained 88.52% accuracy. We also created an Android app to provide the user with an interactive and friendly medium which one can operate, as they won't have to waste time in learning code instead they can directly interact with the app and have the desired result.

REFERENCES

1. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ''International application of a new probability algorithm for the diagnosis of coronary artery disease,'' Amer. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989..
2. C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.298115.
3. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 07562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
4. Hosseinzadeh, F., KayvanJoo, A.H., Ebrahimi, M. et al. Prediction of lung tumor types based on protein attributes by machine learning algorithms. SpringerPlus 2, 238 (2013). https://doi.org/ 10.1186/2193-1801-2-238.
5. R. Thomas and J. E. Judith, "Hybrid Outlier Detection in Healthcare Datasets using DNN and One Class-SVM," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1293-1298, doi: 10.1109/ICECA49313.2020.9297401.
6. C. Gao, Y. Zhang, D. Lo, Y. Shi and J. Huang, "Improving the Machine Learning Prediction Accuracy with Clustering Discretization," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022, pp. 0513-0517, doi: 10.1109/CCWC54503.2022.9720805.
7. El Naqa, I., Murphy, M.J. (2015). What Is Machine Learning?. In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1.
8. Trusculescu, A.A., Manolescu, D., Tudorache, E. et al. Deep learning in interstitial lung disease—how long until daily practice. Eur Radiol 30, 6285–6292 (2020). https://doi.org/10.1007/ s00330-020-06986-4.
9. S. Daberdaku, E. Tavazzi and B. D. Camillo, "Interpolation and K-Nearest Neighbours Combined Imputation for Longitudinal ICU Laboratory Data," 2019 IEEE International Conference on Healthcare Informatics (ICHI), 2019, pp. 1-3, doi: 10.1109/ICHI.2019.8904624.
10. W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," in IEEE Access, vol. 8, pp. 28808-28819, 2020, doi: 10.1109/ACCESS.2019.2955754.
11. D. -E. -M. Nisar, R. Amin, N. -U. -H. Shah, M. A. A. Ghamdi, S. H. Almotiri and M. Alruily, "Healthcare Techniques Through Deep Learning: Issues, Challenges and

Opportunities," in IEEE Access, vol. 9, pp. 98523-98541, 2021, doi: 10.1109/ACCESS.2021.3095312.
12. R. Lee and C. Chitnis, "Improving Health-Care Systems by Disease Prediction," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 726-731, doi: 10.1109/CSCI46756.2018.00145.
13. P. Amudha and S. Sivakumari, "Big data Analytics Using Support Vector Machine," 2018 International Conference on Soft-computing and Network Security (ICSNS), 2018, pp. 1-6, doi: 10.1109/ICSNS.2018.8573641.
14. S. Tayeb et al., "Toward predicting medical conditions using k-nearest neighbors," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3897-3903, doi: 10.1109/BigData.2017.8258395.
15. A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
16. Sunita Agarwala, Kumar, A., Dhara, A.K. et al. Special Convolutional Neural Network for Identification and Positioning of Interstitial Lung Disease Patterns in Computed Tomography Images. Pattern Recognit. Image Anal. 31, 730–738 (2021). https://doi.org/10.1134/ S1054661821040027.
17. A. D. Gunasinghe, A. C. Aponso and H. Thirimanna, "Early Prediction of Lung Diseases," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033668.
18. Angelini, E., Shah, A. Using Artificial Intelligence in Fungal Lung Disease: CPA CT Imaging as an Example. Mycopathologia 186, 733–737 (2021). https://doi.org/10.1007/s11046-021-00546-0.
19. Kwekha-Rashid, A.S., Abduljabbar, H.N. & Alhayani, B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Appl Nanosci (2021). https://doi.org/10.1007/ s13204-021-01868-7.
20. D. S. R. Krishnan, S. C. Gupta and T. Choudhury, "An IoT based Patient Health Monitoring System," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 01-07, doi: 10.1109/ICACCE.2018.8441708

★★★