Smart System to Examine Heart Disease Using Machine Learning Methods

AMAN LAKHMANI[a]*, AYUSH SINGH[a], ATHARV ARYA[a]*****

[a] School of Electronics Engineering (SENSE), Vellore Institute of Technology, Vellore 632014, India

* *e-mail: aman.lakhmani2019@vitstudent.ac.in*

** *e-mail: ayush.singh2019@vitstudent.ac.in*

*** *e-mail: atharv.arya2019@vitstudent.ac.in*

UNDER GUIDANCE OF
POONKHUZALI R
rpoonkuzhali@vit.ac.in

Abstract - Early diagnosis of heart diseases is very important for the patient's life, as it is one of the crucial impacts of mortality in our country. In clinical data analysis, predicting cardiovascular disease is a primary challenge. Recent advances in healthcare technologies have introduced new means of diagnosis of these diseases, one is through Deep Learning (DL). Deep Learning is a type of machine learning that replicates the working of the human brain, it also creates patterns for decision making. These deep learning algorithms are capable of preventing disease outbreaks, detecting and diagnosing diseases, and minimizing ongoing costs for hospital management and patients. This paper proposes a system to avoid these sudden deaths that uses a data set of Heart and to identify these diseases that are categorized into the cardiovascular system, it's predicted through the use of Neural Network with Back Propagation, one of the Deep Learning Techniques throughout the given data sets. In last all these data are used to make a user-friendly interface that can predict the disease by observing the data of concerned patients and informing/warning them whether they got a disease or not and we'll also include a comparison between algorithms that were previously used in this research and challenges faced by deep learning models are discussed.

Key Words - Deep learning, Neural Network, cardiovascular system, algorithms

Introduction

Due to increasing environmental degradation, unhealthy lifestyle, and other factors which is resulting in increasing heart diseases among the elder people as well as young. 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, 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, it can lead to demise of 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 through the net - both thru 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.

By using a dataset for heart disease, 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, this will help an individual to get full information about their health in one place.

Literature Review:

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 Logistic Regression, K-Nearest Neighbor, Artificial Neural Network, Support Vector Machine, and Decision Trees 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 jupytor 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. 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

Methodology:

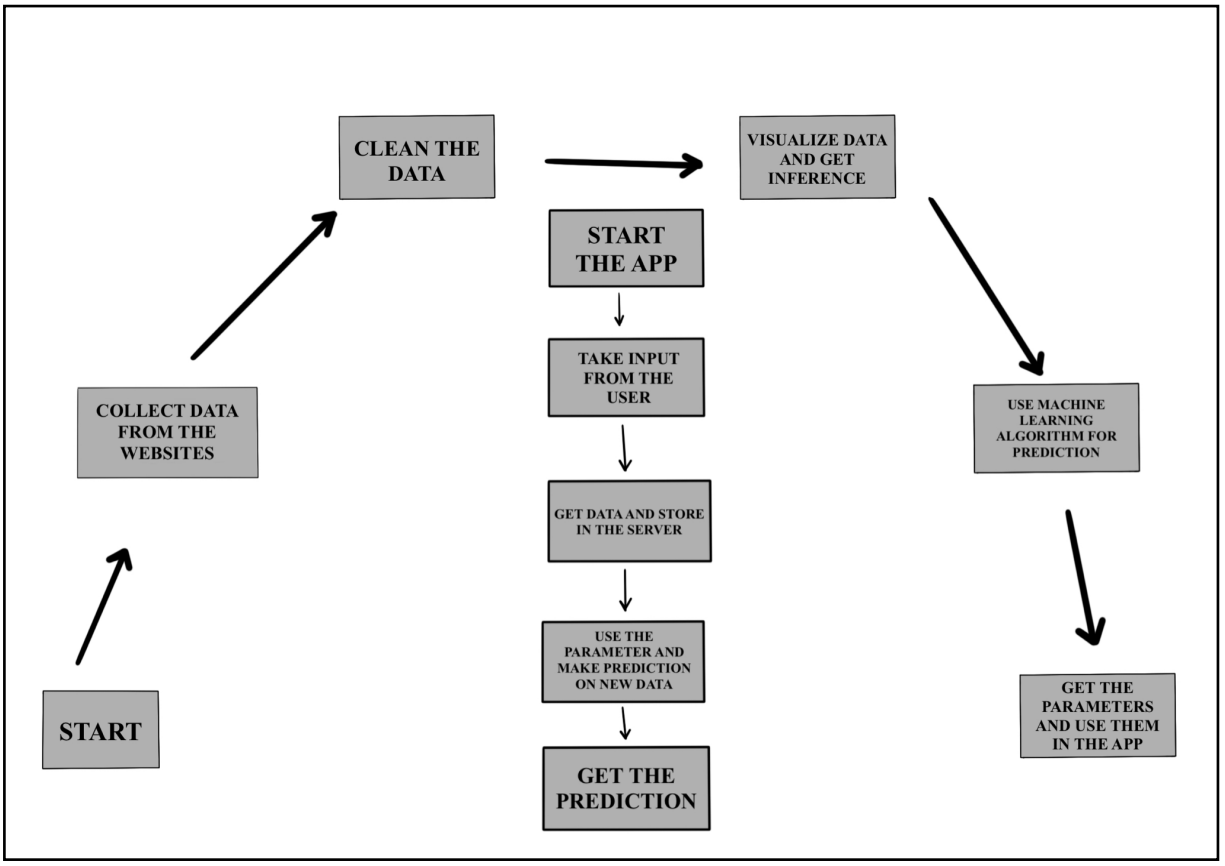

Fig.1 **Proposed methodology of heart disease prediction.**

In Fig1, We are using the UCI dataset and implementing a machine learning algorithm to get hyper-parameters of the suitable algorithm used in the app. The app takes the feature set as input from the user. Based on the parameters of the algorithm the app computes the possibility of getting the disease.

Machine Learning Methods

The K-Nearest-Neighbors (KNN) is a non- parametric classification algorithm, It is a supervised learning algorithm. A training database with a label is provided when data points are divided into different classes, so that a non-labeled data class is predicted. In Classification, the different factors determine the category of unlabeled data that is part of it. KNN is widely used as a separator. It is used to classify data based on training examples in or near a particular area. This method is used to simplify and lower calculation time. The KNN classifier is a nonparametric component that provides good performance for accurate k values. In KNN law, the test sample is the most representative class between training samples close to k, and the distinction is made by calculating the distance between the selected features and the neighbors closest to k (it uses the Euclidean range to calculate its closest neighbors). Major drawbacks to KNN (1) its low efficiency - being a lazy learning method prevents it from many applications such as dynamic web mining of large repositories, and (2) its dependence on selecting "good value" for k.

Random Forest is a meta-learner who works by building multiple cutting trees during training. The

RF method only requires determining the two parameters for creating a guessing model, which includes the number of desired splitting trees and the prediction variability. Simply put, in order to divide a database, a fixed number of random predictions is used, and each sample of the database is divided into a few defined trees.
SVM is an algorithm to find a hyperplane or to make a decision border between two classes predict labels using a vector for one or more features. SVM selects extreme vectors that help create hyperplane. Logistic Regression is a process of modeling the probability of the discrete outcome of a given input variable. A binary outcome is the most common logistic regression model. It is a powerful supervised machine learning algorithm for binary classification. Logistic regression [Eq.1] is a simple and more efficient method for binary and linear classification models. It is easy to realize and has a good performance with linearly separable classes. It is a statistical method for binary classification and can be generalized to multiclass classification.

$$odds = \frac{p}{1-p} \rightarrow \log it\,(p) = \ln\left(\frac{p}{1-p}\right)$$

Let's set logit of P to be equal to $mx + b$, therefore:

$$\log it\,(p) = mx + b \rightarrow mx + b = \ln\left(\frac{p}{1-p}\right)$$

$$\left(\frac{p}{1-p}\right) = e^{(mx+b)} \rightarrow P = \frac{e^{(mx+b)}}{1+e^{(mx+b)}} \rightarrow P\,(x) = \frac{1}{1+e^{-(mx+b)}} \qquad \textbf{.....(1)}$$

This helps logistic regression in having a significant advantage to be used both for classification and class probability. A linear combination of features is taken and a nonlinear sigmoidal function is applied. In the basic version of logistic regression, the output is taken as binary and can be extended to multiple classes.

Results:

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 heart disease are usually 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 is given in:

Accuracy of algorithms for Heart Disease:

```
clf1=LogisticRegression()
clf1.fit(x_train,y_train)
pred1=clf1.predict(x_test)
print(clf1.coef_)
s1=accuracy_score(y_test,pred1)
score.append(s1*100)
print(s1)
```

```
[[-0.02697627 -0.84022786  0.78586918 -0.32852507 -0.30524116  0.02772712
   0.28584229  0.63966602 -0.48408765 -0.44368395  0.28440968 -0.69398945
  -0.47739103]]
0.8852459016393442
```

Fig.2**Accuracy of Logistic regression.**

```
knn = KNeighborsClassifier()
knn.fit(x_train,y_train)

y_true0 = knn.predict(x_test)
s2 = accuracy_score(y_test,y_true0)
score.append(s2*100)
print(s2)
```

```
0.8688524590163934
```

Fig.3 **Accuracy of K-Neighbors**

```
rf = RandomForestClassifier()
rf.fit(x_train,y_train)

y_true1 = rf.predict(x_test)
s4 = accuracy_score(y_test,y_true1)
score.append(s4*100)
print(s4)
```

```
0.8688524590163934
```

Fig.4 **Accuracy of Random Forest**

Fig.5 **Accuracy of SVM.**

Algorithms:	Accuracy (Heart Disease)
Logistic Regression	88.52%
K-Neighbors	86.88%
Random Forest	86.88%
Support Vector Machine	85.24%

Table1**:** Here, algorithms are compared based on their accuracy for heart disease accordingly.

Since not every person is trained enough in using machine learning methods to input the parameters and get the results as shown in Fig 2, we have also created an app that contains the algorithm with the best accuracy. The user only has to enter the parameters associated with him in the user-friendly interface and the already trained algorithm in the app will give the result to him after which he can take further steps to fix an appointment with the doctor.

Data description:

The data includes 303 patient-level features including if they have heart disease at the end or not. Features are like;

Age: Obvious one... Sex: 0: Female 1: Male

Chest Pain Type:

0: Typical Angina
1: Atypical Angina
2: non-Anginal Pain
3: Asymptomatic

Resting Blood Pressure: Person's resting blood pressure.

Cholesterol: Serum Cholesterol in mg/dl

Fasting Blood Sugar:
0: Less Than 120mg/ml
1: Greater Than 120mg/ml

Resting Electrocardiographic Measurement:

0: Normal
1: ST-T Wave Abnormality
2: Left Ventricular Hypertrophy

Max Heart Rate Achieved: Maximum Heart Rate Achieved

Exercise Induced Angina:

1: Yes
0: No

ST Depression: ST depression induced by exercise relative to rest.

Slope: Slope of the peak exercise ST segment:

0: Upsloping
1: Flat
2: Downsloping

Thalassemia: A blood disorder called 'Thalassemia':

0: Normal
1: Fixed Defect
2: Reversible Defect

The number of Major Vessels: Number of major vessels colored by fluoroscopy.

Mobile application:

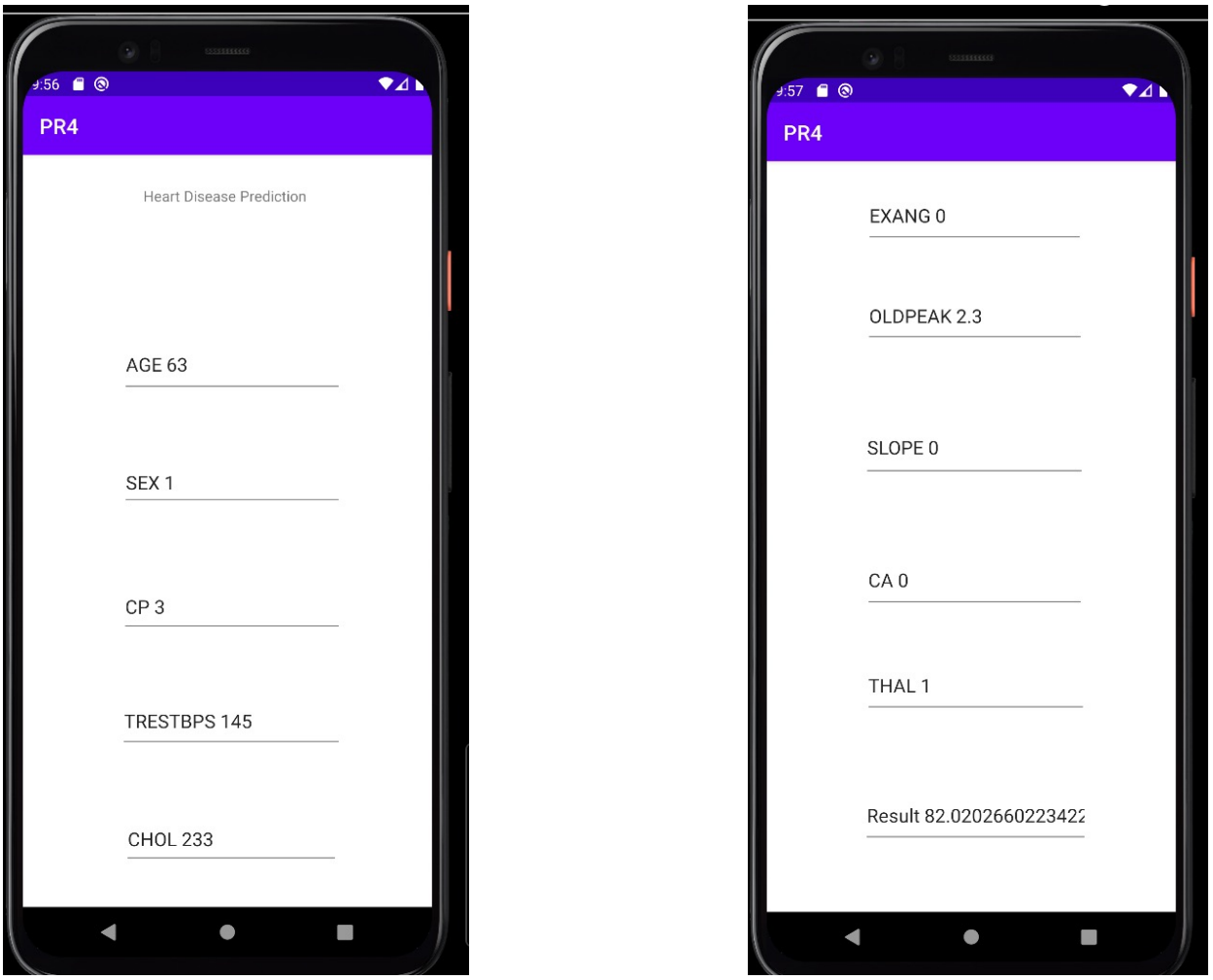

Fig.10; Giving input as parameter in the app.

Fig.10 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 normalising 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 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 as shown in Table1. 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