Multi-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 depicted as conditions that require proceeding with medical consideration, limit everyday exercises, or both. Chronic diseases, for example, heart disease, pneumonia, kidney disease and diabetes are the main sources of death and disability on the planet. It is a critical challenge to detect these diseases by regular clinical data analysis. Early prediction of these chronic diseases would help in saving multiple lives. Machine learning models bring effective solutions for accurate predictions and decision making. Machine learning techniques have shown huge development in the medical industry. The paper attempts to do the predictive analysis of chronic diseases using machine learning. Chronic Kidney Disease will be predicted using Logistic Regression and Random Forest, Diabetes can be predicted using Logistic Regression (LR) and K-Nearest Neighbour (KNN), Heart Disease can be predicted using Random Forest Regression and Decision Tree and finally for Pneumonia can be predicted using a Convolutional Neural Network (CNN) on chest x-rays. An integrated model of the listed algorithms is proposed to predict the listed chronic diseases.***

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

1. INTRODUCTION

Machine Learning is a branch of Artificial Intelligence, where the main goal is to enable the computer to learn from a given set of information. The structure of the data is perceived, 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.

Even though it being a field within computer science, it differs from the standard computational methodologies. Machine Learning calculations allow computers to learn on information data sources and make use of measurable examination to yield esteems that fall inside a particular reach, though in standard computing calculations are sets of expressly modified directions involved by computers for critical thinking or estimations. This methodology in Machine Learning works with computers in building models from test information, to computerize dynamic cycles in light of information 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.

1. 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 Pima Indian Diabetes Database was used by the authors in paper [1], where they evaluated various machine learning algorithms depending on their diabetes detection accuracy. 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%. 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. Random Forest and Decision Tree 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 Random Forest and Logistic Regression. 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 the paper [11] pre-processed the pneumonia dataset from Kaggle, consisting of several X-ray images of patients, by resizing the images and creating a function to store 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%.

1. HEALTHCARE ANALYTICAL MODEL

As we have proposed a system for predicting multiple chronic diseases we have experimented with Diabetes, Chronic Kidney Disease (CKD), Heart Diseases, COVID-19, and Pneumonia. The datasets that are procured are used in accordance with the recommendations and guidance provided by the organizations.

The main objective of our system is to reduce the diagnostic time taken and in turn help the doctor administer treatment at the earliest. To measure the correctness of the models that are used for the prediction we would be computing the Confusion matrix of models, A Confusion matrix consist of true positives, true negatives, false positives, and false negatives,

True Positives are the patients having the disease that are rightly diagnosed by the system. True negatives are the patients that not having the disease that is rightly diagnosed by the system. False Positives are the patients not having the disease that are diagnosed by the system to have the disease. False negative are the patients that have the disease but are wrongly classified as not having the disease. Our attempt is to increase the True positives and True negatives while we are reducing False positives and False negatives to a minimum.

From the values procured from the confusion matrix we can find the Accuracy, Misclassification rate, Recall and Precision of the model, helping us judge and compare the various machine learning models.

1. *Accuracy*

The classifier's accuracy is expressed as the number of samples that were correctly corrected out of a total number of samples, and it is represented as follows:

1. *Misclassification Rate*

It is defined as the classifier's error rate that is the case where classification is not at par with reality.

1. *Recall*

It is defined as the classifier’s predicting the positive values to actual positive values.

1. *Precision*

It is defined as the classifiers rate of predicting the positive

values to the total positive values.

1. PROPOSED SYSTEM & EXPERIMENTAL SETUP

Flowchart diagram of the proposed chronic disease detection system is shown in figure 1.

***A screenshot of a computer

Description automatically generated with low confidence***

Fig. 1 Schematic diagram of the proposed system

1. *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.

1. *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.

1. *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.

1. *Experimental setup*

We have used a different dataset to conduct supervised learning on them to find the best of the classifier for each of the diseases. The attributes of the dataset consist of the medical features that are required to diagnose the chronic disease. We have worked on feature reduction to select the best features for the classification. For Chronic Kidney Disease we have used the UCI Chronic Kidney Disease Dataset following the guidelines provided by UCI. For Diabetes we have used the PIMA Indian diabetes. The UCI heart disease dataset is used for detection of heart disease. The researchers of Qatar University have compiled the COVID-QU-Ex dataset, which consists of 33,920 chest X-ray (CXR) images including: 11,956 COVID-19, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia) and 10,701 Normal images of chest X-rays.

Once the models for detection are trained, we have used Anvil to provide the frontend for the prediction system. Anvil is a platform that can be used for full stack development and deployment of machine learning models with python based back end and web based front end.

1. PROPOSED ALGORITHM
2. *Architecture Diagram*

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

Diagram

Description automatically generated

Fig. 2. Architecture diagram of the proposed system

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.

1. *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. Random sampling is used to train the characteristics for sampling nodes, by combining several decisions to form a single decision. The algorithm consolidates multiple forests on different subsets of a dataset and averages the results to enhance the performance of the dataset’s detection accuracy. The accuracy of the machine can be further improved by combining several classifiers.

**Step 1:** Select 'm' random features from a total of 'n' features, such that m << n.

**Step 2:** Use best split tocalculate node 's' from the selected 'm' features.

**Step 3:** Use best split to further split node ‘s’ into daughter nodes.

**Step 4:** Repeat until the number ‘1’ is reached.

**Step 5:** Repeating the preceding processes ‘n’ number of times, until a forest of 'n' trees is built.

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].

1. *Machine learning algorithm for Diabetes Detection*

**K-Nearest Neighbour:** KNN algorithm arranges another information guide in view of comparability toward accessible information. This infers that new information can be handily arranged into a classification of values utilizing the KNN algorithm. KNN received an accuracy of 79% [1] and 79% [3] in detection of diabetes on PIMA dataset.

**Step 1:** Select a value from the dataset.

**Step 2:** Determine the value of K, such that it is the value of the nearest data point.

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

* Figure the distance between every K and each column of the preparation information.
* Arrange the distances in ascending order.
* Select top n columns from the arranged cluster

Allot a class to the test point in view of 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 predicts the result of a categorically reliant variable, where the result should be a categorical or discrete value. The output will have probabilities of either 0 or 1. Logistic regression is a highly efficient method for linear classification as well as binary classification problems like the problem at hand. 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.

1. *Machine learning algorithm for Heart Disease Detection*

**Decision Tree:** A classification model which is like a tree. It builds a structure of nodes and branches from the evidence collected during the learning phase of the model for each attribute. The connection of the nodes and branches is determined by the number of entities in the dataset. Each attribute has a set number of values used by the forwarding process. Decision for each transaction is reached by following the rules described on each node and branch. At last, the class label will be assigned to the records according to the decision node. This procedure has multiple iterations and will repeat till a class category is assigned to each transaction. In all, attributes are converted to nodes and branches and one of them is selected at the decision.

**Step 1:** Import dependencies like the libraries and dataset

**Step 2:** Split the dataset into a Train set and a Test set for learning and validation.

**Step 3:** Use the train set to train a Decision Tree model

**Step 4:** Use the Test set to predict results using the trained model.

**Step 5:** Compare the Actual Values with Predicted Values

**Step 6:** Visualise the Results of the Decision Tree Regression model.

We are going to use the decision tree for Heart disease detection, as the outputs of 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:** Aggregate of multiple decisions is taken in Random Forest Regression. A complex problem is solved by combining multiple classifiers and that also helps improve the machine's accuracy. Amidst the learning phase a forest is generated which consists of multiple trees. This classifier consolidates multiple forests for different subsets of a dataset and averages out the results to increase the accuracy of the machine's detection capabilities.

**Step 1:** Import dependencies like the libraries and dataset

**Step 2:** Split the dataset into the Train set and Test set for learning and validation.

**Step 3:** Use the train set to train the Random Forest Regression model.

**Step 4:** Use the Test set to predict results using the trained model.

**Step 5:** Compare the Actual Values with Predicted Values

**Step 6:** Visualise 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].

1. *Machine learning algorithm for Pneumonia*

**Convolutional neural network:** The system figures out how to perform feature extraction in a convolutional neural network. The central idea of CNN is to utilize convolution of images and filters to create invariant features which are pushed to the following layer. The qualities of the following layer are joined with various filters to create more invariant and abstract features. This work is done till the last component/yield (say face of X) is accomplished. Convolutional neural network comprises of a few structure components, for example, pooling layers, convolution layers, and completely associated layers, which intended to gain proficiency with the spatial ordered progressions of elements by using a background application.

1. RESULT

While conducting the research we found that the dataset for Chronic Kidney Disease, Heart Disease and Diabetes need feature selection. Due to many attributes present in the dataset results produced by the model would be inconsistent. We have performed pairwise correlation for each dataset attribute with respect to the target variable. Fig 3, Fig 4, and Fig 5 show the bar graph of correlations for each dataset.

Chart, bar chart

Description automatically generated

Fig.3. Correlation between Chronic Kidney Disease and Numeric features

Chart, bar chart

Description automatically generated Fig.4. Correlation between Diabetes and Numeric features

Chart, bar chart

Description automatically generated Fig.5. Correlation between Heart Disease and Numeric features

The accuracies of all chronic disease detection models were assessed and determined using several records that were not a part of the dataset.

It was found that the implemented models for the detection of CKD had an accuracy of 97.5% in both, Random Forest Classifier, as well as Logistic Regression. The resulting accuracies were a statistical average of all the accuracies of our referenced papers. Hyperparameter tuning was performed along the lines of the referenced papers to improve the performance of the models, while also maintaining a higher level of accuracy for the detection of CKD.

The detection of Diabetes was performed using Logistic Regression, as well as KNN, where the implemented models had an accuracy of 81% and 85% respectively. During data pre-processing, it was observed that the number of samples with outcome 0 (Diabetes) were significantly higher than those samples with outcome 1 (No diabetes). Up sampling was performed to balance the dataset with respect to the target variable, while averting the possibility of misclassification of any other attribute. This resulted in getting a higher accuracy than the referenced papers which implemented KNN for their detection model. Hyperparameter tuning was performed to increase the accuracy of detecting Diabetes using Logistic Regression, while a standardscaler function was utilised to standardise the data across the dataset, boosting the average accuracy of the model by 3% as compared to the referenced papers which implemented LR for their detection model.

Decision Tree Classifier & Random Forest Classifier were implemented for the detection of heart disease, we achieved an accuracy of 78.57% and 84.29% respectively. Accuracies achieved by us are above the average accuracies from our referenced papers. Further we performed hyperparameter tuning on the models to raise the accuracy. Post tuning the accuracy of both the models was raised by 2.8%.

We have implemented CNN with transfer learning which has given us an accuracy of 94%. Transfer learning is utilized when the machine stores the information acquired while taking care of one issue and applying it to an alternate yet related problem. In our case the machine has applied its learnings from Pneumonia to Covid 19 detection. Initially we used only CNN which gave us an accuracy of 88% which was boosted by applying transfer learning. We have implemented a heatmap on the output image to show where the layers convolute as shown in Figure 9.

Figure 6, Figure 7 and Figure 8 show the comparisons of accuracy, misclassification rate, recall and precision for each of the models implemented for the diseases respectively.

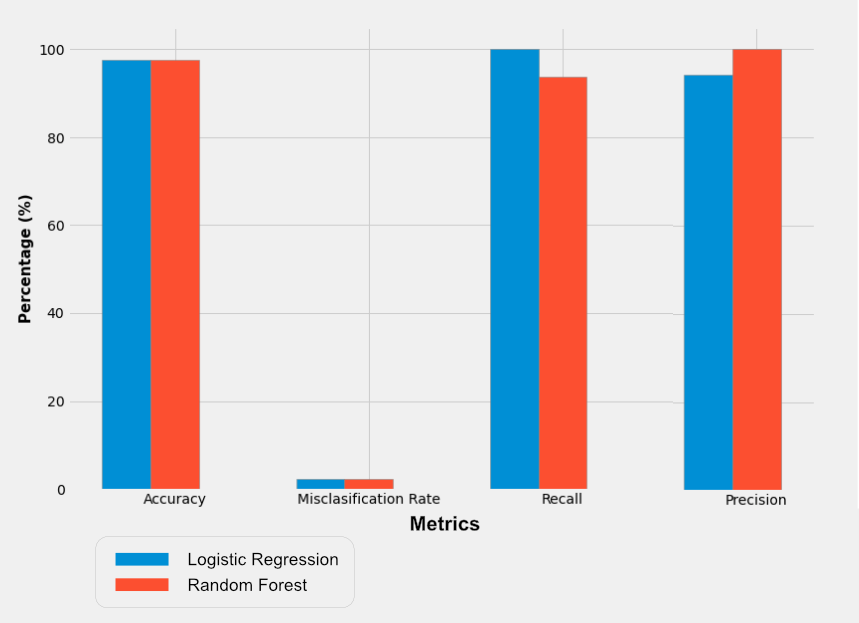


Fig.6. Metrics for models to predict Chronic KidneyDisease

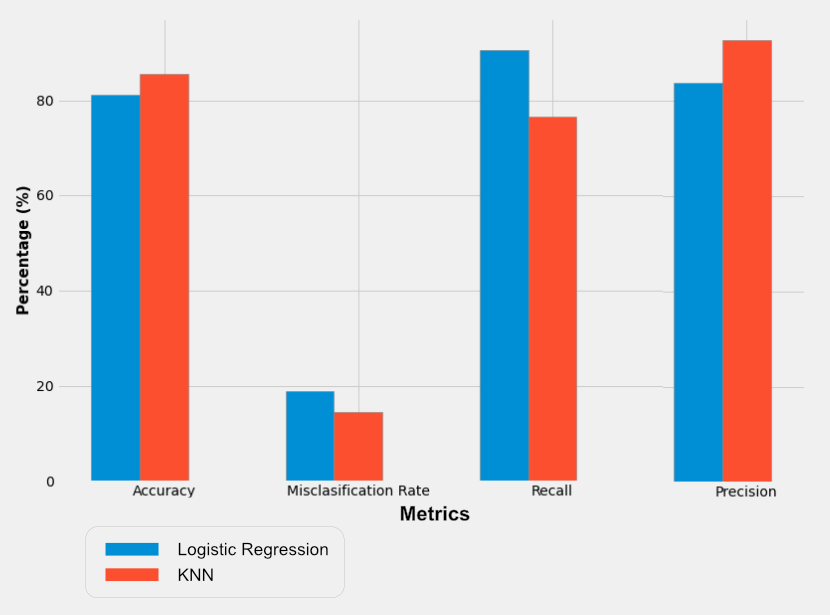


Fig.7. Metrics for models to predict Diabetes

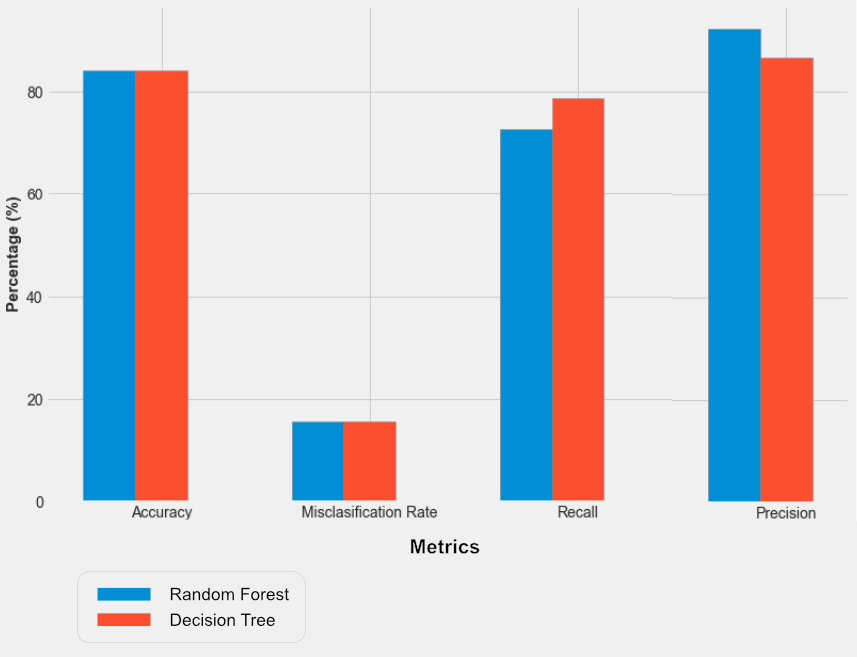


Fig.8. Metrics for models to predict Heart Disease



Fig.9. Heatmap of output image to showing layers convolution

CONCLUSION

Expert systems in medication are characterized as systems with the capacity to catch and store expert information, realities, and thinking strategies to help doctors in diagnosing a patient's condition. These systems endeavour to mirror a specialist's expertise by applying a few computational strategies to help in decision making and critical thinking, by concocting contemplated determinations for a patient's disease or condition. Our project incorporates the core elements of an expert system by supporting medical experts with their claims and help them in early diagnosis of a chronic disease in their patients using trained machine learning models giving high accuracy of disease detection and prediction.

1. 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/1

[12] A. M. Tahir, M. E. H. Chowdhury, A. Khandakar, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M. S. Rahman, S. Al-Madeed, S. Mahmud, M. Ezeddin, K. Hameed, and T. Hamid, “COVID-19 Infection Localization and Severity Grading from Chest X-ray Images”, Computers in Biology and Medicine, vol. 139, p. 105002, 2021,

[13] Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin, “COVID-QU-Ex .” Kaggle,

[14] T. Rahman, A. Khandakar, Y. Qiblawey A. Tahir S. Kiranyaz, S. Abul Kashem, M. Islam, S. Al Maadeed, S. Zughaier, M. Khan, M. Chowdhury, "Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-rays Images," Computers in Biology and Medicine, p. 104319, 2021

[15] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M. E. H. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, and M. Gabbouj, "Covid-19 infection map generation and detection from chest X-ray images," Health Inf Sci Syst 9, 15 (2021),

[16] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, M. T. Islam, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," IEEE Access, vol. 8, pp. 132665-132676, 2020,