**Project on**

**A Detailed Analysis on Kidney and Heart Disease Prediction using Machine Learning**

*a project report submitted in partial fulfillment of the requirement for the award of a degree of*

**Bachelor of Technology**

**in**

**Information Technology**

**SUBMITTED BY:**

**Abhisek Dash 1901106456**

**Samyak Pradhan 1901106520**

**Sarthak Siddhant Bharadwaj 1901106431**

**Swastik Mohapatra 1901106442**

**UNDER THE GUIDANCE OF**

**Ms. Jayshree Dev**

**Faculty, IT**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH BHUBANESWAR, ODISHA**

**ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH BHUBANESWAR, ODISHA**

****

**Certificate**

This is to certify that the project entitled, “**A Detailed Analysis on Kidney and Heart Disease Prediction using Machine Learning**”, is a bonafide work done by “**Abhisek Dash, Swastik Mohapatra, Samyak Pradhan and Sarthak Siddhant Bharadwaj”** in partial fulfillment of requirements for the award of Bachelor of Technology Degree in Information Technology at “**Odisha University of Technology & Research**” is an authentic work carried out by them under my supervision and guidance. The matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree or Diploma to the best of my knowledge.

(Guide)

(Head of the Department)

Date:

Place: OUTR, Bhubaneswar (External Examiner)

**ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH BHUBANESWAR, ODISHA**

****

**Acknowledgement**

We would like to express our sincere regards and profound sense of gratitude to those who have helped us in completing the project successfully. At the very outset, our special and heartfelt thanks to our project supervisor **Ms. Jayshree Dev, Faculty, IT** for her precious guidance, assistance, and constant supervision to bring this piece of work into the present form.

We also extend our humble gratitude to **Dr. Ranjan Kumar Dash, HOD, IT** for providing an opportunity to present this project at the university.

**Abhisek Dash- 1901106456**

**Swastik Mohapatra- 1901106442**

**Samyak Pradahan- 1901106520**

**Sarthak Siddhant Bharadwaj - 1901106431**

**ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH BHUBANESWAR, ODISHA**

****

**Declaration**

We declare that this written submission represents our ideas in our own words and wherever others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the Institute as deemed fit.

**Abhisek Dash- 1901106456**

**Swastik Mohapatra- 1901106442**

**Samyak Pradahan- 1901106520**

**Sarthak Siddhant Bharadwaj - 1901106431**

**Abstract**

Detection of chronic kidney and cardiovascular disease at earlier stages and providing early medical assistance in order to lessen the death rate is the most challenging one. This is due to the absence of proper data for diagnosis, absence of self awareness and the absence of proper technology. A number of dataset on kidney disease and cardiovascular disease are available publicly. The number of features varies from one dataset and to other dataset. We cannot claim that if more is the feature, better is the performance. Dimension reduction is the process of extracting the required features from the given dataset for disease detection. This is necessary for disease prediction with the limited computing resource available with us. A number of classifiers are there for kidney disease and cardiovascular disease detection. Some of them are:- logistic regression, k-nearest neighbor, random forest model etc. Performance of all these classifiers are not same. This may be due to the inherent knowledge in the dataset and the parameters chosen for disease detection. Another problem behind the inaccurate result is that not all classifiers suits to all datasets though the disease is same. Therefore, there is the requirement of better classifier that can predict the disease accurately.

Datasets of different diseases are available online with different numbers of features corresponding to a particular disease. Many dimensionality reduction and feature extraction techniques are used nowadays to reduce the number of features in the dataset and find the most appropriate ones. This paper explores the difference in performance of different machine learning models on the datasets of Chronic kidney disease and Cardiovascular disease. Further, the authors apply Logistic Regression, K Nearest Neighbour, a hybrid model(KNN + LR) and Random Forest Model on the datasets and compare the performance of the models. A key challenge in the field of data mining and machine learning is building accurate and computationally efficient classifiers for medical applications. With an accuracy of 97% in chronic kidney disease and 85% for heart disease, logistic regression classifier and random forest were revealed to be the most optimal method of predictions for kidney and heart disease respectively.

**Signature of Guide Signature of Students** Ms. Jayshree Dev Swastik Mohapatra - 1901106442 Samyak Pradhan - 1901106520

Sarthak S Bharadwaj - 1901106431

Abhisek Dash - 1901106456

**Contents**

**1. Introduction 1 2. Related Works 2 3. Motivation 5 4. Methodology 6**

4.1 Chronic Kidney Disease **6** 4.2 Heart Disease **9 5. Machine Learning Algorithms 12** 5.1 Logistic Regression **12** 5.2 K-Nearest Neighbor **13** 5.1 Random Forest **15 6. Implementation Details 17** 6.1 System Specifications **17** 6.2. Python Libraries **17 7. Code 20** 7.1 Heart Disease **20** 7.2 Chronic Kidney Disease **26 8. Results 33** 8.1. Prediction Model For Chronic Kidney Disease **33** 8.2. Prediction Model For Heart Disease **37 10. Conclusion and Future Works 41 References 42**

**List of tables**

**SL. NO. TABLE NAME PAGE NO**

**1**

Table 2.1: Detailed study of

**4**

previous models

**2**

Table 4.1: Attributes of CKD

**7**

Dataset

**3** Table 4.2: Attributes for Heart diseases dataset

**10**

**4**

Table 8.1: Results on Chronic

kidney disease dataset

**5**

Table 8.2: Results on Heart

disease dataset

**33 37**

**List of Figures**

**SL. NO. FIGURE NAME PAGE NO**

**1**

Figure 4.1 Architecture for

**6**

analysis of CKD

**2**

Figure 4.2 Architecture for

**9**

analysis of heart diseases

**3** Figure 5.1 Sigmoid function of Logistic Regression

**12**

**4**

Figure 5.2 Architecture for

analysis of CKD

**5**

Figure 5.3 Working of the

Random Forest Algorithm

Figure 8.1: Graph of accuracies

**6**

of individual models of Chronic

Kidney Disease

**14 16 34**

**7** Figure 8.2: Confusion Matrix For Knn(CKD**)**

**35**

**12**

**13 14**

**15**

Figure 8.7: Confusion Matrix For Logistic Regression(Heart Disease)

Confusion Matrix For Knn(Heart Disease)

Figure 8.9:Confusion Matrix For Hybrid Model(Heart Disease)

Figure 8.10: Confusion Matrix For Random Forest(Heart Disease)

**38**

**39 39**

**40**

**Abbreviation**

**NKF -** National Kidney Foundation

**CKD** - Chronic Kidney Disease

**CVD** - Cardiovascular Disease

**ICNC** - International Conference on Nuclear Cardiology and Cardiac CT **GFR** - Glomerular filtration rate

**ANFIS**- adaptive neuro-fuzzy inference system

**NN-GA** - Neural Networks using Genetic Algorithms

**PET** - Positron Emission Tomography

**CTI** - Computed Tomography Imaging

**MRI** - Magnetic Resonance Imaging

**SVM** - Support Vector Machine

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**1. INTRODUCTION**

Kidney and heart are the main organs in the human body and require extra care and attention to remain healthy. In this era of modernization where humans are exposed to polluted air, bad lifestyle, consumption of packaged food high in transfat, and more interaction with the electronic gadgets rather than family members, friends and relatives, the prevalence of chronic kidney disease and cardiovascular disease is increasing tremendously. According to a report released in 2019 by National Kidney Foundation (NKF), 10% of the population worldwide is affected by chronic kidney disease (CKD), and it is ranked as the 6th deadliest disease worldwide causing 2.4 million deaths per year. According to World Bank [2] four out of 5 cardiovascular diseases (CVD) related deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. To predict the occurrence of these diseases at earlier stages, machine learning techniques have proved to be a boon to medical practitioners [1].

A study claimed that the modern bedrock of artificial intelligence is machine learning that could predict the occurrence of heart attack with an accuracy of more than 90 percent. The study was presented at The International Conference on Nuclear Cardiology and Cardiac CT (ICNC) 2019. An algorithm 'learned' how imaging data interacts by analyzing 85 variables in 950 patients with known six-year outcomes repeatedly. Identification of patterns was then done correlating the variables to heart attack with an accuracy of more than 90 percent .The use of risk scores is done by doctors to make decisions during treatment. But these scores are based on a subset of variables and generally have moderate accuracy in individual patients [4].

Heart diseases have emerged as one of the most prominent causes of death around the world, claiming 17.7 million lives every year and constituting to 31% of global deaths [5] . Kidney disease is another major health problem contributing to high mortality. Over 1 million people in 112 lower income countries die from kidney failure [6]. Medical organizations around the world collect data on health-related issues to predict the occurrence of a disease at an early stage.

1 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**2. RELATED WORKS**

The human body's two major organs, the kidneys and the heart, need special care and attention to stay healthy. Chronic kidney disease and cardiovascular disease are becoming far more common in this era of modernization, where people are more exposed to air pollution, unhealthy lifestyles, packaged foods high in trans fat, and more interaction with electronic devices than with their families, friends, and loved ones.

The [7] with the aim of accurate prediction of chronic kidney disease progression over time examined the dataset of newly diagnosed patients for over 10 clinical years. Having the threshold value of 15cc/kg/min/1.73 m2 of GFR (Glomerular filtration rate), the author used Takagi-Sugeno type adaptive neuro-fuzzy inference system (ANFIS) model to predict GFR values. Considering the variables like age, weight, sex, diastolic blood pressure, underlying disease, phosphorus, calcium, creatinine, GFR, and uric acid; the prediction model was formulated which despite high uncertainties of human body and dynamic nature of disease progression, predicted the accurate GFR variations.

A study on the dataset of 400 patients (250 CKD patients) wherein 24 attributes were considered for analysis [8]. Using the classification algorithm K-nearest neighbors, neural networks, and random forest; and feature reduction methods i.e. wrapper and LASSO regularization method, the analysis was done. Results of the analysis reveal that random forest algorithm with 12 attributes can detect CKD with accuracy of 99.8% using the F1-measure model and 0.107 root mean square error.

The [9] used common spatial pattern and linear discriminant analysis for identifying the attributes that dominantly contribute in chronic kidney disease. Having the most accurate result determination by common spatial pattern, the analysis worked on classifying the dataset into non-CKD and CKD. The analysis revealed that specific gravity, albumin, hypertension, hemoglobin, and diabetes are the most important attributes of determining CKD.

2 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

The [10] performed classification of the dataset into non-chronic and chronic kidney disease using support vector machine neural networks. The study of 400 patients revealed that about 94.44% of accuracy is derived by having 6 attributes. The work [11] focused on the chronic kidney disease prevalence in the United States by using the predictive analytics techniques like Logistic Regression, Random trees, Artificial Neural networks, Chi-square automatic interaction detector, support vector machines, and Naïve Bayes for predicting the chronic kidney disease. Herein, having the preprocessing of data and imputing missing data, the analysis shows that random trees provide most accurate information for CKD prevalence.

In [5] identified the appropriate diet plan for CKD patients by using classification algorithms. Using the predictive potassium zone, the experiment was performed via algorithms like multiclass decision forest, multiclass decision jungle, multiclass logistic regression, and multiclass neural network. The analysis with 99.17% accuracy revealed that multiclass decision forest is the most effective algorithm for identifying a diet plan for CKD patients.

The [6] added more information to chronic kidney diseases by proposing a genetic algorithm trained neural network-based model (NN-GA) for assessing the newest threats of this disease in underdeveloped and developing nations. This proposed model was compared to the multilayer perception feed forward network, random forest, and neural networks which revealed that in terms of accuracy, recall, F-measure, and precision NN-GA model is more efficient than other existing models.

The [7] studied 224 chronic kidney disease patients’ records on UCI machine learning repository. Predicting the model based on deep neural networks, the absence of presence of CKD could be predicted by 97% accuracy. Reducing the chances of overfitting, the built in model provided better results compared to any other algorithms.

In [2] classifier, the dataset of 400 patients and their 25 attributes has been examined. The analysis by the authors on the recall, precision, F1-score, and specificity revealed that with the

3 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

accuracy of 100%, multimodal model has performed better than other conventional classifiers. The [1] for the CKD patients proposed a hybrid intelligent model by using logistic regression and neural network techniques. Having the accuracy of 97.8%, the analysis revealed that hybrid intelligent model was superior to models by 64%.

**Table 2.1: Detailed study of previous models**

| **Sl. no.** | **Journal detail and Title** | **Tools/Algorithms**  **Advantage Limitation**  **used** |
| --- | --- | --- |
| 1 | Salekin and Stankovic [12] | K-NN, RF, and  Detection F1-score  Small dataset size with  NN,  of RF 99.8  missing values was  Wrapper approach  used;  and  Severity level  Embedded  prediction  approach  was not included |
| 2 | Priyanka et al. [13] | NB, KNN, SVM,  NB, KNN, SVM,  Small size dataset.  DT, and  DT, and ANN.  No stages prediction.  ANN. NB  NB accuracy is  Feature extraction was  94.6%  not carried out and  classification  accuracy needs  improvement |
| 3 | Yashfi [14] | RF and ANN RF and ANN with  Small size dataset and  an accuracy  no  of 97.12% and  stages prediction  94.5% |
| 4 | Rubini PE[15] | RF, LR, SVM, NB RF, LR, SVM, NB  Small size dataset,  with an accuracy of  Lower accuracy with  84.81%, 83.82%  Naive Bayes  ,74.05% & 54.08% |
| 5 | Sahoo[16] | SVM, LR, KNN SVM, LR, KNN  Small size dataset  with an accuracy  85.2% |

4 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**3. MOTIVATION**

Machine learning techniques have become a hotspot in biomedical and health-care research with the development of electronic health records and medical information. Using informatics technology in medical, healthcare can significantly alter and subvert the conventional healthcare and medical services [2]. In order to eventually enhance human health, new models and methods for early diagnosis of illness, care, and prevention are being conceived [3]. Machine learning is the modern bedrock of artificial intelligence. Through repetition and adjustment, it is capable of exploiting huge amounts of information and determine complicated patterns which are otherwise time-consuming for humans. This study aims to improve upon the existing models and tackle their limitations by using a different combination of algorithms to build a more accurate and reliable model for predicting the likelihood of chronic kidney disease (CKD) and heart diseases.

5 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

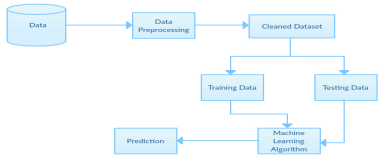
Machine Learning

**4. METHODOLOGY**

Initially the dataset was collected from an open-source dataset platform. The initial step involved data pre-processing i.e. examination of the nature of data via mean and median in order to fill in missing data or have the normality assessment in the dataset. Further, in the second phase the analysis commenced with data-classification. Herein, the entire dataset was randomly divided into two categories i.e. training data and testing data (80% and 20%). Then, a soft learning algorithm (Logistic Regression, KNN and Random forest) was built for the training data in order to have the identification of classification technique. Lastly, evaluating the accuracy, precision, recall, F1-measure, and support with the testing data based on the soft learning algorithm, the results were drawn. Hence, providing the final information about the performance of all models for the heart disease and kidney disease, the analysis helped address the aim of the research.

The dataset selected for the analysis consist of 400 chronic kidney disease patients’ data and 4240 heart disease patients’ data.

**4.1. CHRONIC KIDNEY DISEASE**

**Fig 4.1: Architecture for analysis of CKD**

6 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**4.1.1. DATASET**

Dataset of prediction of chronic kidney disease using machine learning algorithm is downloaded from Kaggle repository. In that dataset 400 patient records are included. It contains 25 attributes but only 14 attributes are taken for building the model. Age, Blood pressure, Albumin, Red blood cells, Pus cell, Pus cells clumps, Serum creatinine, Haemoglobin, White blood cell count, Red blood cell count, Anaemia, Classification, Appetite, Packed cell volume all these 14 attributes are used to build the model.

**4.1.2. DATA PREPROCESSING**

Data Reduction: Out of 25 attributes present in the dataset, 14 important attributes have been selected that are required to build the predictive model. Following table shows the selected attributes:

**Table 4.1: Attributes of CKD Dataset**

| **Attribute** | **Value Used** |
| --- | --- |
| Age | Discrete Integer Values |
| Blood Pressure | Discrete Integer Values |
| Albumin | Nominal Values |
| Red blood cells | Nominal Values(Normal, Abnormal) |
| Pus cells | Nominal Values(Normal, Abnormal) |
| Pus cells clumps | Nominal Values(Present, Not-Present) |

Serum creatinine Numeric Values

Haemoglobin Numeric Values

7 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

| White blood cell count | Discrete Integer Values |
| --- | --- |
| Red blood cell count | Numeric Values |
| Anemia | Nominal Values(Yes, No) |
| Classification | Nominal Values(CKD, Not CKD) |
| Appetite | Nominal Values(Good, Poor) |
| Packed cell volume | Discrete Integer Values |

Data Cleaning: Open source raw data of CKD patients available on the internet is gathered from Kaggle. Missing values in the dataset like NA’s or blank values are removed by using the simple imputer function which uses the “most frequent” (mode of the data) as its strategy parameter. Categorical data is converted numerical data using label encoder and fit\_transform function. Two rows of redundant data are removed from the dataset. So, the final dataset contains 398 records.

**4.1.3. TRAINING AND TESTING DATASET**

The dataset is divided into two sub datasets both containing 14 attributes.

Training data: Training dataset is derived from the main dataset and it contains 318 out of 398 records in the main dataset of CKD.

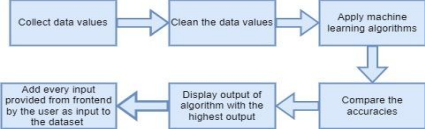
Testing data: Testing dataset is 80 out of 398 records from the main CKD dataset. 8 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**4.2. HEART DISEASE**

.

**Figure 4.2: Architecture for analysis of heart diseases**

**4.2.1. DATASET**

The dataset which was used for analysis is “Framingham” obtained from Kaggle. For the heart disease, the database is of 4240 patients represent 16 attributes i.e. age, sex, education, current smoking status, cigarettes smoke per day, BP meds, prevalent stroke, prevalent Hypertension, diabetes, total cholesterol, sys BP, diaBP, BMI, heart rate, glucose, and ten year CHD(result) . The dataset is obtained from Kaggle.

**4.2.2.DATA PRE-PROCESSING**

Data is cleaned by replacing all the non-available values of the columns with uneven distributions across a range of values with the median value and for those with even distributions with the mean value of the data in the column.Categorical data are assigned with numerical values. After treating the missing values, one row is dropped from the dataset. So, the final dataset contains 4239 records.

9 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Table 4.2: Attributes for Heart diseases dataset**

| **Attribute** | **Value Used** |
| --- | --- |
| Age | Discrete Integer Values |
| Sex | Nominal Values(Male, Female) |
| Education | Nominal Values |
| Current smoking status | Nominal Values(Yes, No) |
| Cigarettes smoke per day | Discrete Integer Values |
| BP meds | Nominal Values(Present, Not-Present) |
| Prevalent smoke | Nominal Values(High, Low) |
| Prevalent Hypertension | Nominal Values(Yes, No) |
| Diabetes | Nominal Values(Present, Not-present) |
| Total cholesterol | Numeric Values |
| sys BP | Numeric Values |
| diaBP | Numeric Values |
| BMI | Numeric Values |
| Heart rate | Numeric Values |
| Glucose | Numeric Values |
| Ten year CHD | Nominal Values |

**\*** The nominal values are represented as integer values in the dataset.

**4.2.3. TRAINING AND TEST DATASET**

The dataset is divided into two sub datasets both containing 16 attributes.

Training data: Training dataset is derived from the main dataset and it contains 3391 out of 4239 records in the main dataset of heart disease.

10 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

Testing data: Testing dataset contains 848 out of 4239 records in the main dataset of heart disease.

11 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

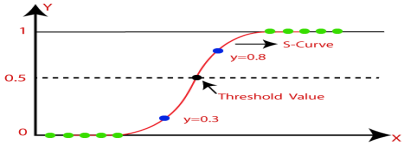
Machine Learning

**5. MACHINE LEARNING ALGORITHMS**

**5.1. LOGISTIC REGRESSION**

Logistic regression is a machine learning model targeting the binary classification. Working in a similar form as linear regression, this method could be used for providing information about the dataset and building in the relationship between one dependent variable and other variables. The probability of a target variable is predicted using the supervised learning classification algorithm known as logistic regression. Since the dependent variable's nature is dichotomous, there are only two viable classes. Simply said, the dependent variable is a binary variable, with data recorded as either 1 (which represents success/yes) or 0 (which represents failure/no).

A logistic regression model makes mathematical predictions about P(Y=1) as a function of X. One of the most basic machine learning algorithms, it may be applied to a number of categorization issues, including spam identification, diabetes prediction, cancer diagnosis, etc. Predicting the probability, threshold value could be derived. Herein, considering t as linear function in univariate regression model, the logistic equation is expressed as per equation (1). 

The threshold value defined by the model could be affected by the precision or recall values. **Figure 5.1: Sigmoid function of Logistic Regression**

12 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

The following presumptions concerning logistic regression must be understood before we begin its implementation:

● When using binary logistic regression, the outcome of interest is represented by factor level 1 and the target variables are required to always be binary.

● The independent variables in the model must be independent of one another in order to prevent multicollinearity.

● Our model must contain relevant variables.

● For logistic regression, a high sample size is recommended.

Logistic regression is categorised into the following groups according to the number of categories:

1. Binomial: The target variable can only be one of two potential types, such as "0" or "1," which can indicate various outcomes such as "win" or "loss," "pass" or "fail," "dead" or "alive," etc.

2. Multinomial: The target variable may include three or more distinct types that are not ordered and lack quantitative significance, such as "illness A" vs. "disease B" vs. "disease C."

3. Ordinal: Using ordered categories as the target variables, ordinal analysis is used. A test score can be classified as "extremely poor," "poor," "good," or "very good," for instance. Here, a score of 0, 1, 2, or 3 can be assigned to each category.

**5.2. K- NEAREST NEIGHBOR**

The KNN method is the simplest algorithm for prediction using the Euclidean distance, thus also called lazy learning. K defines the neighbor number by considering nearest neighbor votes for prediction.

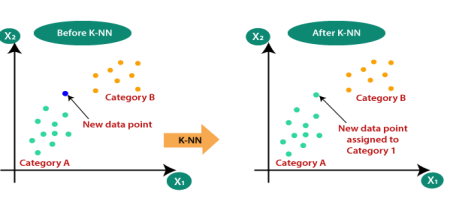
13 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

The KNN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.A new data point is classified using the KNN algorithm based on similarity after all the existing data has been stored. This means that utilising the KNN method, fresh data can be quickly and accurately sorted into a suitable category.Although the KNN approach is most frequently employed for classification problems, it can also be utilised for regression.Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

The KNN algorithm simply stores the dataset during the training phase, and when it receives new data, it categorises it into a category that is very similar to the new data.

**Figure 5.2: KNN Algorithm demonstration**

The following algorithm can be used to describe how the KNN works:

Step 1: Decide on the neighbors’ K-numbers.

Step 2: Calculate the Euclidean distance between K neighbors in step two.

Step 3: Based on the determined Euclidean distance, select the K closest neighbors. Step 4: Count the number of data points in each category among these k neighbors. Step 5: Assign the fresh data points to the category where the neighbor count is highest. Step six: Our model is complete.

14 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

The case wherein x and y define independent attribute, D defines their distance and K the nearest neighbor, output value could be predicted as per equation (2):



Although the good value of K is selected via heuristic techniques, presence of irrelevant or noisy features could degrade the accuracy of this method.

**5.3. RANDOM FOREST**

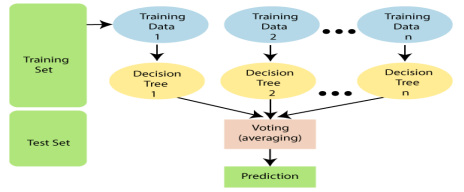
Being an ensemble learning algorithm, random forest (random decision forest) is used for handling classification and regression problems. Suitable for the dataset wherein the decision tree is used for training, this method uses bootstrap aggregation technique. Thus, instead of having output derivation by just one decision tree, the method works with combined decision trees. As while splitting nodes, this method helps in the best or most vital feature, thus good and effective results could be derived using this algorithm in data mining [1].

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

Higher accuracy and overfitting are prevented by the larger number of trees in the forest. 15 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 5.3 : Working of the Random Forest algorithm**

Some decision trees may predict the correct output, while others may not, because the random forest combines numerous trees to forecast the class of the dataset. But when all the trees are combined, they forecast the right result. Consequently, the following two presumptions for an improved Random forest classifier:

● For the dataset's feature variable to predict true outcomes rather than a speculated result, there should be some actual values in the dataset.

● Each tree's predictions must have extremely low correlations.

First, N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase.

The stages can be used to demonstrate the working process:

**Step 1**: Pick K data points at random from the training set.

**Step 2**: Construct the decision trees linked to the chosen data points (Subsets). **Step 3**: Select N for the size of the decision trees you wish to construct.

**Step 4**: Repeat steps 1 and 2 in step 4.

**Step 5**: Assign new data points to the category that receives the majority of votes by looking up each decision tree's predictions for the new data points.

16 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**6. IMPLEMENTATION DETAILS**

**6.1. SYSTEM SPECIFICATIONS**

For testing the model, the following configurations have been used:

● Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz 2.40 GHz processor

● 16 GB of memory with Windows 64-bit OS (version 11)

The code for the model has been written using Google Colab( a python web ide). **6.2. PYTHON LIBRARIES**

**6.2.1. Numpy**

● NumPy (Numerical Python) is a Python library that is used for scientific computing and data analysis.

● NumPy allows for the creation and manipulation of multi-dimensional arrays, which are particularly useful for representing and analyzing complex data structures.

● NumPy provides a wide range of mathematical functions that can be applied to arrays, including functions for linear algebra, Fourier analysis, and statistical analysis.

**6.2.2. Pandas**

● Pandas is a popular open-source library for data manipulation and analysis in Python. It provides easy-to-use data structures and data analysis tools for handling structured data, such as CSV files, Excel spreadsheets, SQL databases, and JSON files.

● The core data structure in Pandas is the DataFrame, which is a 2-dimensional table-like data structure consisting of rows and columns, similar to a spreadsheet or SQL table. The library also provides the Series data structure, which is a one-dimensional labeled array capable of holding any data type.

17 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

● Pandas offers powerful functionalities for data manipulation, including selecting, filtering, sorting, grouping, and merging data. It also supports advanced data cleaning and transformation operations, such as missing data handling, reshaping, and pivoting.

**6.2.3. sklearn**

● scikit-learn, also known as sklearn, is a popular Python library for machine learning. It provides a range of tools for tasks such as classification, regression, clustering, and dimensionality reduction. It is built on top of other popular scientific libraries in Python, such as NumPy, SciPy, and matplotlib.

● sklearn provides a consistent API for all its models, making it easy to switch between models and algorithms.

● sklearn includes a large selection of models for various tasks, such as linear and logistic regression, support vector machines, decision trees, and random forests.

● sklearn provides a range of evaluation metrics to assess the performance of models, such as accuracy, precision, recall, and F1 score.

**6.2.4**. **streamlit**

● It provides an easy-to-use interface that allows developers to quickly create interactive dashboards, data visualizations, and web applications without the need for web development skills.

● Streamlit can be installed using pip, and requires only a few lines of code to get started. ● Streamlit allows developers to easily share and deploy their applications to the web, making it easy to collaborate with others and showcase their work.

**6.2.5. Pickle**

● Pickle is a Python module used for object serialization and deserialization. Serialization is the process of converting an object in memory to a byte stream that can be stored on disk or transmitted over a network. Deserialization is the reverse process of restoring the object from its serialized form.

18 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

● Pickle can be used to store and retrieve objects in a binary format, which is useful for saving the state of an application or for transmitting objects over a network.

19 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**7.CODE**

**7.1. HEART DISEASE**

7.1.1.IMPORTING DEPENDENCIES

**import numpy as np**

**import pandas as pd**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.neighbors import KNeighborsClassifier**

**from sklearn.metrics import accuracy\_score**

**from sklearn.metrics import classification\_report**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**import pickle**

**from sklearn.metrics import confusion\_matrix**

**from sklearn.metrics import ConfusionMatrixDisplay**

7.1.2. DATA COLLECTION AND PREPROCESSING

**# loading csv to pandas dataframe**

**data = pd.read\_csv('/content/framingham.csv')**

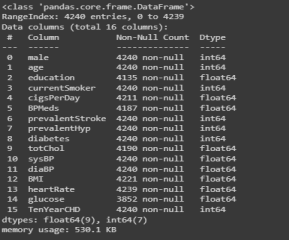
**# getting some info about dataset**

**data.info()**

20 |Page

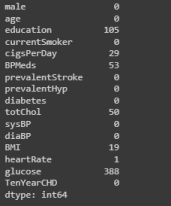
A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning



**# checking for missing values**

**data.isnull().sum()**

****

**# filling missing values in education with the mean**

**data['education'].fillna(data['education'].mean(), inplace=True)** 21 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**# filling missing values in other columns with mean or median data['cigsPerDay'].fillna(data['cigsPerDay'].median(), inplace=True) data['BPMeds'].fillna(data['BPMeds'].median(), inplace=True) data['totChol'].fillna(data['totChol'].mean(), inplace=True) data['BMI'].fillna(data['BMI'].mean(), inplace=True)**

**data['glucose'].fillna(data['glucose'].mean(), inplace=True)**

**# removing row with missing heartRate value**

**data.dropna(subset=['heartRate'], inplace=True)**

**# checking the distribution of target variable**

**data['TenYearCHD'].value\_counts()**

****

****

7.1.3.SPLITTING THE FEATURES AND TARGET

**X = data.drop(columns='TenYearCHD', axis=1)**

**Y = data['TenYearCHD']**

7.1.4. SPLITTING THE DATA INTO TRAINING AND TEST DATA

22 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=0)**

****

**Y\_train.value\_counts()**

****

**Y\_test.value\_counts()**

****

7.1.5. MODEL TRAINING

7.1.5.1..Logistic Regression

**lr = LogisticRegression()**

**# training the LogisticRegression model with Training data lr.fit(X\_train, Y\_train)**

**print(classification\_report(Y\_test, X\_test\_pred\_lr))**

****

23 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

7.1.5.2. K nearest neighbours

**knn = KNeighborsClassifier()**

**# training the KNN model with training data**

**knn.fit(X\_train, Y\_train)**

**print(classification\_report(Y\_test, X\_test\_pred\_knn))**

****

7.1.5.3. Hybrid model(Logistic regression + KNN)

**estimators = [('lr', lr), ('knn', knn)]**

**ensemble = VotingClassifier(estimators, voting='soft')**

**ensemble.fit(X\_train, Y\_train)**

**pred = ensemble.predict(X\_test)**

**accuracy = accuracy\_score(Y\_test, pred)**

**print("Accuracy:", accuracy)**

****

**print(classification\_report(Y\_test, pred))**

****

24 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

7.1.5.4. Random Forest

**from sklearn.ensemble import RandomForestClassifier**

**max\_accuracy = 0**

**for x in range(200):**

**rf = RandomForestClassifier(random\_state=x)**

**rf.fit(X\_train,Y\_train)**

**X\_test\_pred\_rf = rf.predict(X\_test)**

**current\_accuracy = round(accuracy\_score(X\_test\_pred\_rf, Y\_test)\*100,2) if(current\_accuracy>max\_accuracy):**

**max\_accuracy = current\_accuracy**

**best\_x = x**

**print(max\_accuracy)**

**print(best\_x)**

**rf = RandomForestClassifier(random\_state=best\_x)**

**rf.fit(X\_train,Y\_train)**

**X\_test\_pred\_rf = rf.predict(X\_test)**

****

**# Accuracy on test data**

**test\_accuracy\_rf = accuracy\_score(X\_test\_pred\_rf, Y\_test) print('Accuracy on test data :', test\_accuracy\_rf)**

****

**print(classification\_report(Y\_test, X\_test\_pred\_rf))**

25 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning



**7.2. CHRONIC KIDNEY DISEASE**

7.2.1**.**IMPORTING THE LIBRARIES

**import warnings**

**warnings.filterwarnings("ignore")**

**import numpy as np**

**import pandas as pd**

**from sklearn.neighbors import KNeighborsClassifier**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.preprocessing import MinMaxScaler**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.metrics import accuracy\_score**

**from sklearn.metrics import classification\_report**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**import pickle**

**from sklearn.metrics import confusion\_matrix**

**from sklearn.metrics import ConfusionMatrixDisplay**

7.2.2. DATA COLLECTION AND PROCESSING

**# loading a csv data to a pandas dataframe**

**kidney\_data = pd.read\_csv('/content/kidney\_disease.csv')**

**kidney\_data.info()**

26 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning



7.2.3. DELETING THE REDUNDANT ATTRIBUTE

**kidney\_data.drop('id', inplace=True, axis=1)**

**kidney\_data.drop('sg', inplace=True, axis=1)**

**kidney\_data.drop('su', inplace=True, axis=1)**

**kidney\_data.drop('ba', inplace=True, axis=1)**

**kidney\_data.drop('bgr', inplace=True, axis=1)**

**kidney\_data.drop('bu', inplace=True, axis=1)**

**kidney\_data.drop('sod', inplace=True, axis=1)**

**kidney\_data.drop('pot', inplace=True, axis=1)**

**kidney\_data.drop('htn', inplace=True, axis=1)**

**kidney\_data.drop('dm', inplace=True, axis=1)**

**kidney\_data.drop('cad', inplace=True, axis=1)**

**kidney\_data.drop('pe', inplace=True, axis=1)**

7.2.4.CHECKING THE NULL VALUES

**kidney\_data.isnull().sum()**

27 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning



7.2.5.IMPUTING THE NULL VALUES

**from sklearn.impute import SimpleImputer**

**imp\_mode = SimpleImputer(missing\_values=np.nan,strategy ='most\_frequent')**

**kidney\_data\_imputed=pd.DataFrame(imp\_mode.fit\_transform(kidney\_data)) kidney\_data\_imputed.columns=kidney\_data.columns**

**kidney\_data\_imputed**

7.2.6. LABEL ENCODING OF CATEGORICAL VALUES TO NUMERIC VALUES **kidney\_data\_imputed.drop(index=[37,230],axis=1,inplace=True) from sklearn.preprocessing import LabelEncoder**

**le = LabelEncoder()**

**label = le.fit\_transform(kidney\_data\_imputed['rbc'])**

**kidney\_data\_imputed.drop("rbc", axis=1, inplace=True)**

**kidney\_data\_imputed["rbc"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['pc'])**

**kidney\_data\_imputed.drop("pc", axis=1, inplace=True)**

**kidney\_data\_imputed["pc"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['pcc'])**

**kidney\_data\_imputed.drop("pcc", axis=1, inplace=True)**

**kidney\_data\_imputed["pcc"] = label**

28 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**label = le.fit\_transform(kidney\_data\_imputed['pcv'])**

**kidney\_data\_imputed.drop("pcv", axis=1, inplace=True)**

**kidney\_data\_imputed["pcv"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['wc'])**

**kidney\_data\_imputed.drop("wc", axis=1, inplace=True)**

**kidney\_data\_imputed["wc"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['rc'])**

**kidney\_data\_imputed.drop("rc", axis=1, inplace=True)**

**kidney\_data\_imputed["rc"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['appet'])**

**kidney\_data\_imputed.drop("appet", axis=1, inplace=True)**

**kidney\_data\_imputed["appet"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['ane'])**

**kidney\_data\_imputed.drop("ane", axis=1, inplace=True)**

**kidney\_data\_imputed["ane"] = label**

**label = le.fit\_transform(kidney\_data\_imputed['classification']) kidney\_data\_imputed.drop("classification", axis=1, inplace=True) kidney\_data\_imputed["classification"] = label**

7.2.7.SPLITTING THE FEATURES AND TARGET

**X = kidney\_data\_imputed.drop(columns='classification',axis=1) Y = kidney\_data\_imputed['classification']**

7.2.8.SPLITTING INTO TRAINING AND TEST DATA

29 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(X,Y,test\_size=0.2,stratify= Y,random\_state=2)**

7.2.9. MODEL TRAINING

7.2.9.1.K Nearest neighbors

**knn=KNeighborsClassifier()**

**# training the KNN model with Training Data**

**knn.fit(X\_train, Y\_train)**

**(X\_test\_pred\_knn, Y\_test)**

**print(classification\_report(Y\_test, X\_test\_pred\_knn))**

****

7.2.9.2.Logistic Regression

**lr = LogisticRegression()**

**# training Logistic Regression model with training data**

**lr.fit(X\_train, Y\_train)**

**print(classification\_report(Y\_test, X\_test\_pred\_lr))**

30 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning



7.2.9.3.Hybrid model(combining Linear Regression and knn)

**estimators = [('knn', knn), ('lr', lr)]**

**ensemble = VotingClassifier(estimators, voting='soft')**

**ensemble.fit(X\_train, Y\_train)**

**pred = ensemble.predict(X\_test)**

**accuracy = accuracy\_score(Y\_test, pred)**

**print("Accuracy:", accuracy)**

****

**print(classification\_report(Y\_test, pred))**

****

7.2.9.4.Random forest

**from sklearn.ensemble import RandomForestClassifier**

**max\_accuracy = 0**

**for x in range(200):**

**rf = RandomForestClassifier(random\_state=x)**

31 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**rf.fit(X\_train,Y\_train)**

**X\_test\_pred\_rf = rf.predict(X\_test)**

**current\_accuracy = round(accuracy\_score(X\_test\_pred\_rf, Y\_test)\*100,2) if(current\_accuracy>max\_accuracy):**

**max\_accuracy = current\_accuracy**

**best\_x = x**

**print(max\_accuracy)**

**print(best\_x)**

**rf = RandomForestClassifier(random\_state=best\_x)**

**rf.fit(X\_train,Y\_train)**

**X\_test\_pred\_rf = rf.predict(X\_test)**

****

**print(classification\_report(Y\_test, X\_test\_pred\_rf))**

****

32 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**8. RESULT**

With the filtering of dataset and having the pre-processing to fill out the missing the data and understand the nature of dataset, final analysis was done to compare the efficiency of models. Finding the most effective model to forecast chronic renal disease and heart disease was made possible even with the comparison of the soft-learning algorithm in the subsection. The analysis, which is based on performance measurements, is displayed in the subsections below.

**8.1. PREDICTION MODEL FOR CHRONIC KIDNEY DISEASE**

**Table 8.1: Results on Chronic kidney disease dataset**

| **ML Models** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Support** |
| --- | --- | --- | --- | --- | --- |
| Logistic  Regression | 97 | 0.98 | 0.97 | 0.98 | 80 |
| KNN | 89 | 0.89 | 0.89 | 0.89 | 80 |
| Random  Forest | 95 | 0.95 | 0.95 | 0.95 | 80 |
| LR + KNN | 94 | 0.94 | 0.94 | 0.94 | 80 |

KNN has 89% accuracy, Logistic Regression accuracy is 97%, the hybrid model(combination of LR and KNN) has 94% accuracy and the random forest accuracy level is 95%. With respect to the precision level, recall level, and F1-score too the situation is similar i.e. Logistic regression algorithm has highest values with precision of 0.98, recall of 0.97, and F1-score of 0.98 while KNN has lowest values with precision of 0.89, recall of 0.89, and F1-score of 0.89. Among all algorithms Random Forest and the hybrid model is at moderate to slightly higher level in their efficiency with precision, recall and F1-score . The support value for all the algorithms is same i.e. of 80.

33 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.1: Graph of accuracies of individual models of Chronic Kidney Disease**

The performance of the classification models for a certain set of test data is evaluated using a matrix called the confusion matrix. Only after the true values of the test data are known can it be determined. Although the matrix itself is simple to understand, some of the terminology used in connection with it might be. It is also referred to as an error matrix since it displays the errors in the model performance as a matrix.

34 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.2: Confusion Matrix For Knn(CKD)**

**Figure 8.3: Confusion Matrix For Logistic Regression(CKD)**

35 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.4:Confusion Matrix For Hybrid Model(CKD)**

**Figure 8.5:Confusion Matrix For Random Forest(CKD)**

36 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**8.2. PREDICTION MODEL FOR HEART DISEASE**

**Table 8.2: Results on Heart disease dataset**

| **ML Models** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Support** |
| --- | --- | --- | --- | --- | --- |
| Logistic  Regression | 84 | 0.80 | 0.84 | 0.78 | 848 |
| KNN | 82 | 0.77 | 0.82 | 0.78 | 848 |
| Random  Forest | 85 | 0.83 | 0.85 | 0.80 | 848 |
| LR + KNN | 84 | 0.79 | 0.84 | 0.78 | 848 |

KNN has 82% accuracy, Logistic Regression accuracy is 84%, the hybrid model(combination of LR and KNN) has 84% accuracy and the random forest accuracy level is 85%. With respect to the precision level, recall level, and F1-score too the situation is similar i.e. Random Forest algorithm has highest values with precision of 0.83, recall of 0.85, and F1-score of 0.80 while KNN has lowest values with precision of 0.77, recall of 0.82, and F1-score of 0.78. Among all algorithms Logistic Regression and the hybrid model are at moderate level in their efficiency with precision, recall and F1-score . The support value for all the algorithms is the same i.e. of 848.

37 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.6: Graph of accuracies of individual models of Heart Disease**

**Figure 8.7: Confusion Matrix For Logistic Regression(Heart Disease)**

38 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.8: Confusion Matrix For Knn(Heart Disease)**

**Figure 8.9:Confusion Matrix For Hybrid Model(Heart Disease)**

39 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**Figure 8.10: Confusion Matrix For Random Forest(Heart Disease)**

40 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**9. CONCLUSION AND FUTURE WORKS**

The discussed work implements different techniques viz. LR, KNN, RF and hybrid model combined with LR and KNN Classifier, which can be used to predict the possibility of occurrence of kidney or heart disease. It is also concluded that logistic regression proves to be a better classifier for kidney disease and random forest for heart disease datasets. Similar prediction systems can be built by calculating the correlation between the discussed diseases and other diseases.

However, there are still some challenges that need to be addressed in the future. One major challenge is the lack of standardization in medical data collection, storage, and analysis. This can lead to inconsistencies and errors in the data, which can affect the accuracy of the prediction models.

Another challenge is the interpretability of the machine learning models. It is important to understand how the models are making predictions and to ensure that they are not biased or discriminatory towards certain groups of patients.

Future work in this area should focus on developing more accurate and interpretable machine learning models that can be used in clinical practice. Additionally, efforts should be made to standardize medical data collection and analysis, as well as to address issues of bias and discrimination in the models. Overall, the use of machine learning in predicting heart and kidney diseases has great potential to improve patient outcomes and reduce healthcare costs.

41 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

**REFERENCES**

[1]. M. E. Farooqui and D. J. Ahmad, “A Detailed Review On Disease Prediction Models That Uses Machine Learning,” International Journal of Innovative Research in Computer Science & Technology, vol. 8, no. 4, Jul. 2020.

[2]. P. M. M. Bhajibhakare, “Heart Disease Prediction using Machine Learning,” International Journal for Research in Applied Science and Engineering Technology, vol. 7, no. 12, pp. 455–460, Dec. 2019.

[3]. S. M. Pasha, “Diabetes and Heart Disease Prediction Using Machine Learning Algorithms,” International Journal of Emerging Trends in Engineering Research, vol. 8, no. 7, pp. 3247–3252, Jul. 2020.

[4]. Dokare, A. Prithiani, H. Ochani, S. Kanjan, and D. Tarachandani, “Prediction of Having a Heart Disease Using Machine Learning,” SSRN Electronic Journal, 2020.

[5]. Singh, “Prediction of Heart Disease using Machine Learning,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 150–166, Jul. 2020.

[6]. K. K. Y., Dr. Anurag Sharma, Dr. Abhishek Badholia, “Heart Disease Prediction Using Machine Learning Techniques,” Information Technology In Industry, vol. 9, no. 1, pp. 207–214, Feb. 2021.

[7]. Dokare, A. Prithiani, H. Ochani, S. Kanjan, and D. Tarachandani, “Prediction of Having a Heart Disease Using Machine Learning,” SSRN Electronic Journal, 2020.

[8]. P. K. Sahoo and P. Jeripothula, “Heart Failure Prediction Using Machine Learning Techniques,” SSRN Electronic Journal, 2020.

[9]. K. Howe, N. James, P. Gladding, C. Prabhakar, A. Gavin, and L. Dawson, “Predicting CRT Response Using Machine Learning Analysis of Pre-Implant ECG Data,” Heart, Lung and Circulation, vol. 26, p. S188, 2017.

42 |Page

A Detailed Analysis on Kidney and Heart Disease Prediction using

Machine Learning

[10]. R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang, and N. Mehendale, “Disease Prediction From Various Symptoms Using Machine Learning,” SSRN Electronic Journal, 2020

[11]. M. S. Devi, P. Swathi, S. S. Upadhyay, N. K. Sah, A. Budhia, vamsi chowdary, S. Srivastava, and M. Rohella, “Feature Predominance Ensemble Inquisition towards Liver Disease Prediction using Machine Learning,” SSRN Electronic Journal, 2021.

[12]. Salekin, Asif, and John Stankovic. "Detection of chronic kidney disease and selecting important predictive attributes." In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 262-270. IEEE, 2016.

[13] Priyanka K, Science BC. Chronic kidney disease prediction based on naive Bayes technique. 2019. p. 1653–9.

[14] Yashfi, Shanila Yunus, Md Ashikul Islam, Nazmus Sakib, Tanzila Islam, Mohammad Shahbaaz, and Sadaf Salman Pantho. "Risk prediction of chronic kidney disease using machine learning algorithms." In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5. IEEE, 2020.

[15] Rubini, P. E., C. A. Subasini, A. Vanitha Katharine, V. Kumaresan, S. Gowdham Kumar, and T. M. Nithya. "A cardiovascular disease prediction using machine learning algorithms." *Annals of the Romanian Society for Cell Biology* (2021): 904-912.

[16] Sahoo, Prasanta Kumar, and Pravalika Jeripothula. "Heart Failure Prediction Using Machine Learning Techniques." *Available at SSRN 3759562* (2020).

[17] Tekale, Siddheshwar, Pranjal Shingavi, Sukanya Wandhekar, and Ankit Chatorikar. "Prediction of chronic kidney disease using machine learning algorithm." *International Journal of Advanced Research in Computer and Communication Engineering* 7, no. 10 (2018): 92-96.

43 |Page