



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Final Year Project Report is the bonafide work of MARY SEREN RANJANA X (40130138), SHERLIE L (40130217) who have done the project work as an individual who carried out the project entitled "HEART AND KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS" under my supervision from November 2023 to April 2024.

Internal Guide

Dr. M.A. MUTHIAH, M.Tech, Ph.D.

Head of the Department

Dr. T. RAVI, M.E., Ph.D.

Submitted for Viva voce Examination held on 18/04/2024

Internal Examiner

External Examiner

DECLARATION

We, **MARY SEREN RANJANA X** (40130138) and **SHERLIE L** (40130217) hereby declare that the Project Report entitled "**HEART AND KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS**" done by us under the guidance of **Dr. M.A. MUTHIAH, M.Tech, Ph.D.**, Assistant Professor, Department of Electronics and Communication Engineering, at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering.

DATE: 18/04/2024

PLACE: Chennai

1. Mary Seren Ranjana X

2. Sherlie L

SIGNATURE OF THE CANDIDATES

Mary Seren Ranjana X

Sherlie L

HEART AND KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

Submitted in partial fulfilment of requirements for the award of
Bachelor of Engineering in Electronics and Communication Engineering

by

MARY SEREN RANJANA X (40130138)

SHERLIE L (40130217)



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SCHOOL OF ELECTRICAL AND ELECTRONICS**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

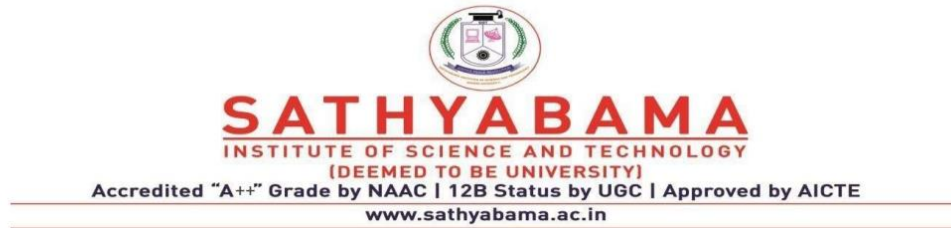
CATEGORY -1 UNIVERSITY BY UGC

Accredited with Grade “A++” by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI-600119

APRIL – 2024



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Final Year Project Report is the bonafide work of MARY SEREN RANJANA X (40130138), SHERLIE L (40130217) who have done the project work as an individual who carried out the project entitled "HEART AND KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS" under my supervision from November 2023 to April 2024.

Internal Guide

Dr. M.A. MUTHIAH, M.Tech, Ph.D.

Head of the Department

Dr. T. RAVI, M.E., Ph.D.

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

We, **MARY SEREN RANJANA X** (40130138) and **SHERLIE L** (40130217) hereby declare that the Project Report entitled “**HEART AND KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS**” done by us under the guidance of **Dr. M.A. MUTHIAH, M.Tech, Ph.D.**, Assistant Professor, Department of Electronics and Communication Engineering, at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering.

1.

DATE:

2.

PLACE:

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey my thanks to **Dr. N. M. NANDHITHA, M.E., Ph.D., Professor & Dean, School of Electrical and Electronics** and **Dr. T. RAVI, M.E., Ph.D., Professor & Head, Department of Electronics and Communication Engineering** for providing us necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr. M.A MUTHIAH, M.Tech, Ph.D., Assistant Professor, Department of Electronics and Communication Engineering**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of Electronics and Communication Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

Cardiovascular diseases (CVDs) and kidney diseases are leading causes of morbidity and mortality worldwide. Early detection and accurate prediction of these diseases are critical for effective intervention and prevention strategies. Machine learning (ML) and deep learning (DL) techniques have shown promising results in various medical applications, including disease prediction. In this study, we propose a novel approach for predicting heart and kidney diseases using ML and DL techniques. First we preprocess the medical data, which includes clinical features, biomarkers, and patient demographics. Next, we apply ML algorithms such as logistic regression, support vector machines, random forests, and gradient boosting machines to build predictive models based on these features. Additionally, we explore DL techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to capture complex patterns and temporal dependencies in the data. We evaluate the performance of each model using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Furthermore, we conduct feature importance analysis to identify the most relevant factors contributing to disease prediction. Our experimental results demonstrate that both ML and DL models achieve high accuracy in predicting heart and kidney diseases. DL models, particularly CNNs and RNNs, outperform traditional ML algorithms, indicating the efficacy of deep learning in medical data analysis. Additionally, feature importance analysis reveals key biomarkers and clinical indicators associated with disease risk. Overall, our study highlights the potential of ML and DL techniques in early detection and prediction of heart and kidney diseases, thereby aiding clinicians in making informed decisions and implementing timely interventions to improve patient outcomes.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	v
	LIST OF ABBREVIATIONS	ix
	LIST OF FIGURES	xi
	LIST OF TABLES	xiii
1	INTRODUCTION	1
	OVERVIEW OF MACHINE LEARNING AND DEEP LEARNING	1
2	LITERATURE SURVEY	3
3	AIM AND SCOPE OF THE EXISTING SYSTEM	7
	3.1 AIM OF THE PROJECT	7
	3.2 SCOPE OF THE PROJECT	8
	3.3 EXISTING SYSTEM	
	3.3.1 SYSTEM ARCHITECTURE FOR EXISTING SYSTEM	9
	3.3.2 EXISTING SYSTEM WORKING	10
	3.4 CLASSIFIERS USED IN EXISTING SYSTEM	11
	3.4.1 K-NEAREST NEIGHBORS	11
	3.4.2 SUPPORT VECTOR MACHINE	12
	3.4.3 NAÏVE BAYES	13
	3.4.4 RANDOM FOREST	14
	3.4.5 DECISION TREE	15
	3.4.6 LOGISTIC REGRESSION	16
	3.4.7 CATBOOST ALGORITHM	17
	3.4.8 XGBOOST ALGORITHM	19
	3.5 MEDICAL DIAGNOSIS	20
	3.6 ALGORITHM EXPLANATION	21

	3.6.1 AR(p) MODELS	21
4	PROPOSED SYSTEM METHODOLOGY	23
	4.1 SYSTEM REQUIREMENTS	23
	4.2 SYSTEM ARCHITECTURE	23
	4.3 PROPOSED SYSTEM	23
	4.4 PROPOSED SYSTEM ARCHITECTURE	24
	4.5 PROPOSED SYSTEM METHODOLOGY	25
	4.6 SYSTEM IMPLEMENTATION	26
	4.6.1 COLLECTING DATASET	26
	4.6.2 PRE- PROCESSING	27
	4.6.3 DATA INPUT	27
	4.7 CONVOLUTIONAL NEURAL NETWORKS	28
	4.7.1 CONVOLUTIONAL NEURAL NETWORKS STEPS INVOLVED IN FUNCTIONING	29
	4.8 COLLECTED DATASET	30
	4.8.1 HEART DATASET	30
	4.8.2 KIDNEY DATASET	31
	4.8.2.1 KIDNEY TUMOR DATASET	31
	4.8.2.2 KIDNEY STONE DATASET	32
	4.8.2.3 KIDNEY CYST DATASET	33
	4.8.2.4 NORMAL KIDNEY DATASET	33
	4.9 WEBSITE CREATED FOR THE PREDICTION PROCESS	34
	4.10 STANDARD	35
	4.11 CONSTRAINT	35
	4.12 TRADE OFF	36
5	RESULT AND DISCUSSIONS	38
	5.1 DATA INPUT PAGE FOR HEART DISEASE PREDICTION	39
	5.2 RESULT OUTPUT PAGE FOR HEART DISEASE PREDICTION	40
	5.3 RESULT OUTPUT PAGES FOR KIDNEY DISEASE	39

	PREDICTION	
	5.3.1 KIDNEY TUMORE PREDICTION	39
	5.3.2 KIDNEY STONE PREDICTION	40
	5.3.3 KIDNEY CYST PREDICTION	40
	5.3.4 NORMAL KIDNEY PREDICTION	41
	5.4 CONFUSION MATRIX	42
	5.4.1 CONFUSION MATRIX OBTAINED FOR HEART	42
	DISEASE PREDICTION	
	5.4.2 CONFUSION MATRIX OBTAINED FOR KIDNEY	42
	DISEASE PREDICTION	
	5.5 TABULATION REPRESENTING IMPORTANT	43
	PARAMETERS FOR PREDICTION	
6	CONCLUSION AND FUTURE WORK	44
	6.1 CONCLUSION	44
	6.2 FUTURE WORK	45
	REFERENCE	46

LIST OF ABBREVIATIONS

ML	MACHINE LEARNING
DL	DEEP LEARNING
CNN	CONVOLUTIONAL NEURAL NETWORKS
RNN	RECURRENT NEURAL NETWORKS
DT	DECISION TREE
NB	NAÏVE BAYES
MLP	MULTILAYER PERCEPTRON
K-NN	K- NEAREST NEIGHBOR
SCRL	SINGLE CONJUNCTIVE RULE LEARNER
RBF	RADIAL BASIS FUNCTION
SVM	SUPPORT VECTOR MACHINE
IoT	INTERNET OF THINGS
CKD	CHRONIC KIDNEY DISEASE
CVD	CARDIO VASCULAR DISEASE
MCC	MATTHEWS CORRELATION COEFFICIENT
DBN	DEEP BELIEF NETWORK
AHDCNN	ADAPTIVE HYBRIDISED DEEP CONVOLUTIONAL NEURAL NETWORKS
SVR	SUPPORT VECTOR REGRESSION
PCA	PRINCIPLE COMPONENT ANALYSIS
IE	INFORMATION EXTRACTION
EOD	END OF DAY
LDA	LINEAR DISCRIMINANT ANALYSIS
BM	BOLTZMAN MACHINE
DDPG	DEEP DETERMINANT POLICY GRADIENT
AR	AUTO-REGRESSIVE
CT	COMPUTED TOMOGRAPHY
MRI	MAGNETIC RESONANCE IMAGING

ANN	ARITHMETIC NEURAL NETWORKS
DNN	DEEP NEURAL NETWORK

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
3.1	REPRESENTATION OF HEART AND KIDNEY IN A HUMAN BODY	7
3.2	EXISTING SYSTEM ARCHITECTURE FOR HEART DISEASE PREDICTION	9
3.3	K- NEAREST NEIGHBORS	11
3.4	SUPPORT VECTOR MACHINE	12
3.5	NAÏVE BAYES	13
3.6	RANDOM FOREST	14
3.7	DECISION TREE	15
3.8	LOGISTIC REGRESSION	16
3.9	CATBOOST ALGORITHM	17
3.10	XGBOOST ALGORITHM	19
4.1	DIAGRAM REPRESENTING PROPOSED SYSTEM ARCHITECTURE	24
4.2	PARAMETERS FOR KIDNEY DISEASE PREDICTION	31
4.3	CT SCAN DATASET COLLECTED FOR KIDNEY TUMOR	32
4.4	CT SCAN DATASET COLLECTED FOR KIDNEY STONE	32
4.5	CT SCAN DATSET COLLECTED FOR KIDNEY CYST	33
4.6	CT SCAN DATSET COLLECTED FOR NORMAL KIDNEY	33
4.7	WEBSITE CRFEATED FOR PREDICTION	34
5.1	DATA INPUT PAGE	38
5.2	HEART DISEASE PREDICTION OUTPUT PAGE	39
5.3	KIDNEY TUMOR PREDICTION	39
5.4	KIDNEY STONE PREDICTION	40
5.5	KIDNEY CYST PREDICTION	40

5.6	NORMAL KIDNEY PREDICTION	41
5.7	RANDOM FOREST CONFUSION MATRIX	42
5.8	SUPPORT VECTOR MACHINE CONFUSION MATRIX	42

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
4.1	TABULATION OF DATASET COLLECTED FOR HEART	30
5.1	DEPICTING THE SENSITIVITY, SPECIFICITY AND ACCURACY	43

CHAPTER - 1

INTRODUCTION

Cardiovascular diseases and kidney diseases are among the leading causes of mortality and morbidity globally, posing significant challenges to healthcare systems worldwide. Early detection and accurate prediction of these diseases are crucial for effective management and timely intervention, ultimately improving patient outcomes and reducing healthcare burdens. Traditional diagnostic approaches often rely on clinical assessment and biomarker analysis, which may lack precision and fail to capture subtle patterns in complex medical data. Recent advancements in machine learning (ML) and deep learning (DL) techniques have revolutionized various domains, including healthcare, by enabling the analysis of large-scale medical data and extraction of valuable insights for disease prediction and diagnosis. ML algorithms, such as logistic regression, support vector machines, random forests, and gradient boosting machines, offer efficient ways to model complex relationships between clinical features and disease outcomes. Meanwhile, DL techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excel in capturing intricate patterns and temporal dependencies within sequential medical data.

1.1 OVERVIEW OF MACHINE LEARNING AND DEEP LEARNING

In this context, this study aims to leverage ML and DL techniques for the prediction of heart and kidney diseases. By analyzing a comprehensive set of clinical features, biomarkers, and patient demographics, we seek to develop accurate and robust predictive models that can aid healthcare providers in identifying individuals at high risk of developing these conditions. Furthermore, by evaluating the performance of both ML and DL models and conducting feature importance analysis, we aim to uncover key factors contributing to disease risk and provide valuable insights for clinical decision-making. The utilization of ML and DL techniques in disease prediction holds

immense potential for enhancing preventive healthcare strategies, enabling early intervention, and improving patient outcomes. By harnessing the power of advanced computational methods, we can move towards personalized medicine approaches that cater to individual patient needs and contribute to the advancement of healthcare delivery systems. This study represents a significant step towards harnessing the potential of data-driven approaches in tackling the challenges posed by heart and kidney diseases, ultimately paving the way for more effective and efficient healthcare solutions.

CHAPTER - 2

LITERATURE SURVEY

Y.S. Sun, et.al [1]Predicting Heart Disease at Early Stages using Machine Learning: A Survey Foreseeing and identification of coronary illness has consistently been a basic and testing task for medical care specialists. Emergency clinics and different facilities are offering costly treatments and activities to treat heart sicknesses. In this way, anticipating coronary illness at the beginning phases will be valuable to individuals around the globe with the goal that they will make important moves prior to getting serious.

Y. Khourdifi , et.al [2]Application of Machine Learning in Disease Prediction The use of AI in the field of clinical determination is expanding step by step. This can be contributed fundamentally to the improvement in the grouping and acknowledgment frameworks utilized in infection determination which can give information that guides clinical specialists in early discovery of deadly sicknesses and consequently, increment the endurance pace of patients essentially.

Y.Lu, et.al [3]A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease The Cleveland informational collection for heart infections, containing 303 examples, has been utilized as the primary data set for the preparation and testing of the created framework. 10-Fold Cross-Validation has been applied to expand the measure of information, which would some way or another have been restricted. Various classifiers, in particular Decision Tree (DT), Naïve Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM), have been utilized.

F.K. Ahmad, et.al [4]Preliminary design of estimation heart disease by using machine learning ANN within one year In this paper talked about the improvement of coronary illness expectation utilizing AI (for this situation the

Artificial Neural Network or ANN). There are 13 factors that can decide coronary illness as per Miss Chaitrali paper. Forecast of an individual's coronary illness one year ahead is performed by considering the model pulse information. Information is taken by utilizing device, for example, keen mirror, brilliant mouse, advanced cells and savvy seat. Pulse information were gathered through the Internet and gathered in a worker. Learning in this framework is performed for a time of one year to get sufficient information to make expectations.

R.Hou, et.al [5]Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques Coronary illness is perhaps the main sources of mortality on the planet today. Expectation of cardiovascular sickness is a basic test in the zone of clinical information examination. AI (ML) has been demonstrated to be successful in helping with settling on choices and expectations from the huge amount of information delivered by the medical care industry. We have additionally seen ML strategies being utilized in ongoing advancements in various regions of the Internet of Things (IoT). Different investigations give just a brief look into foreseeing coronary illness with ML methods.

A.N.Muiru, et.al [6]A Machine Learning Analysis of Health Records of Patients With Chronic Kidney Disease at Risk of Cardiovascular Disease Chronic kidney disease (CKD) describes a long-term decline in kidney function and has many causes. It affects hundreds of millions of people worldwide every year. It can have a strong negative impact on patients, especially when combined with cardiovascular disease (CVD): patients with both conditions have lower survival chances. In this context, computational intelligence applied to electronic health records can provide insights to physicians that can help them make better decisions about prognoses or therapies. In this study we applied machine learning to medical records of patients with CKD and CVD. First, we predicted if patients develop severe CKD, both including and excluding information about the year it occurred or date of the last visit. Our methods achieved top mean Matthews correlation

coefficient (MCC) of +0.499 in the former case and a mean MCC of +0.469 in the latter case.

C.P.Wen, et.al [7]Early Prediction of Chronic Kidney Disease Using Deep Belief Network Chronic kidney disease (CKD) is still a health concern despite advances in surgical care and treatment. CKD's growth in recent years has gained much interest from researchers around the world in developing high-performance methods for diagnosis, treatment and preventive therapy. Improved performance can be accomplished by learning the features that are in the concern of the problem. In addition to the clinical examination, analysis of the medical data for the patients can help the health care partners to predict the disease in early stage. Although there are many tries to build intelligent systems to predict the CKD by analysis the health data, the performance of these systems still need enhancement. This Paper proposes an intelligent classification and prediction model. It uses modified Deep Belief Network (DBN) as classification algorithm to predict the kidney related diseases and the Softmax as activation function and the Categorical Cross-entropy as a loss function. The evaluation of the proposed model shows that the model can predict the CKD with accuracy 98.5%

M.A Hossain, et.al [8]Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform Kidney disease is a severe lifelong condition caused either by renal disease or by impaired functions of the kidneys. In the present area of research, Kidney cancer is one of the deadliest and crucial importance for the survival of the patients diagnosis and classification. Early diagnosis and proper therapy can stop or delay the development of this chronic disease into the final stage where dialysis or renal transplantation is the only way of saving the life of the patient. The development of automated tools to accurately identify subtypes of kidney cancer is, therefore, an urgent challenge in the recent past. In this paper, to examine the ability of various deep learning methods an Adaptive hybridized Deep Convolutional Neural

Network (AHDCNN) has been proposed for the early detection of Kidney disease efficiently and effectively. Classification technology efficiency depends on the role of the data set. To enhance the accuracy of the classification system by reducing the feature dimension an algorithm model has been developed using CNN.

M.A.Hossain, et.al [9]A Deep-Learning-Based System for Automated Sensing of Chronic Kidney Disease In this article, we propose a new sensing technique for the automated detection of kidney disease. The salivary urea concentration is monitored to detect the disease. A new sensing approach is introduced to monitor the urea levels in the saliva sample. Furthermore, for analyzing the raw signals obtained from the sensor, we have implemented a one-dimensional deep learning convolutional neural network (CNN) algorithm, which is incorporated with a support vector machine (SVM) classifier. The use of CNN-SVM integrated network enhanced the classification accuracy of the model. The proposed model successfully classified the samples with an accuracy of 98.04%.

O.J.Wouters, et.al [10]Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening Objective: Chronic kidney disease (CKD) is a major public health concern worldwide. High costs of late-stage diagnosis and insufficient testing facilities can contribute to high morbidity and mortality rates in CKD patients, particularly in less developed countries. Thus, early diagnosis aided by vital parameter analytics using affordable computer-aided diagnosis could not only reduce diagnosis costs but improve patient management and outcomes. Methods: In this study, we developed machine learning models using selective key pathological categories to identify clinical test attributes that will aid in accurate early diagnosis of CKD. Such an approach will save time and costs for diagnostic screening.

CHAPTER – 3

AIM AND SCOPE OF EXISTING SYSTEM

3.1 AIM OF THE PROJECT

The aim of the project is to establish a way to predict heart and kidney diseases in a person using our software application which uses machine learning algorithms and also convolutional neural networks, the classifier which produces the highest accuracy is taken into consideration for predicting if the person is likely to have the disease or not.



Fig: 3.1: Representation of heart and kidney in a human body

3.2 SCOPE OF THE PROJECT

The scope of utilizing machine learning (ML) and deep learning (DL) techniques for heart and kidney disease prediction is vast and holds immense potential in improving healthcare outcomes. Firstly, these techniques offer the opportunity to develop accurate and reliable predictive models by leveraging large-scale datasets comprising clinical features, biomarkers, and patient demographics. By incorporating diverse data sources, including electronic health records, medical imaging, and genetic information, the models can capture complex relationships and patterns that traditional diagnostic methods may overlook. Furthermore, the application in disease prediction enables the identification of high-risk individuals at an early stage, facilitating timely interventions and preventive measures. These models can assist healthcare providers in risk stratification, guiding personalized treatment strategies tailored to individual patient needs. Additionally, these techniques can aid in the discovery of novel biomarkers and risk factors associated with heart and kidney diseases, thereby advancing our understanding of disease pathophysiology and informing future research directions.

Moreover, the scalability and adaptability of ML and DL algorithms make them suitable for integration into existing healthcare systems, offering the potential for widespread adoption and implementation in clinical practice. From screening and diagnosis to prognosis and treatment optimization, these techniques have the capability to revolutionize the entire continuum of care for heart and kidney diseases. In summary, the scope of employing techniques for heart and kidney disease prediction is multifaceted, encompassing improved predictive accuracy, early risk stratification, discovery of novel biomarkers, and seamless integration into clinical workflows. By harnessing the power of data-driven approaches, we can transform the landscape of preventive healthcare and pave the way for more effective management and mitigation of these debilitating conditions.

3.3 EXISTING SYSTEM

The existing system for heart disease prediction using machine learning techniques encompasses various approaches and methodologies aimed at leveraging computational methods to improve predictive accuracy and clinical decision-making. Traditional diagnostic methods often rely on manual assessment of clinical features and biomarkers, which may be time-consuming and subject to interobserver variability. In contrast, ML techniques offer automated and data-driven approaches to analyze large-scale medical data and extract valuable insights for disease prediction. One prevalent approach in the existing system involves the utilization of ML algorithms such as logistic regression, support vector machines, and random forests to develop predictive models based on clinical features, biomarkers, and patient demographics. These models can effectively identify patterns and relationships within the data, enabling the prediction of disease outcomes with high accuracy. Additionally, ensemble methods and feature selection techniques are employed to enhance model performance and interpretability.

3.3.1 SYSTEM ARCHITECTURE FOR EXISTING SYSTEM

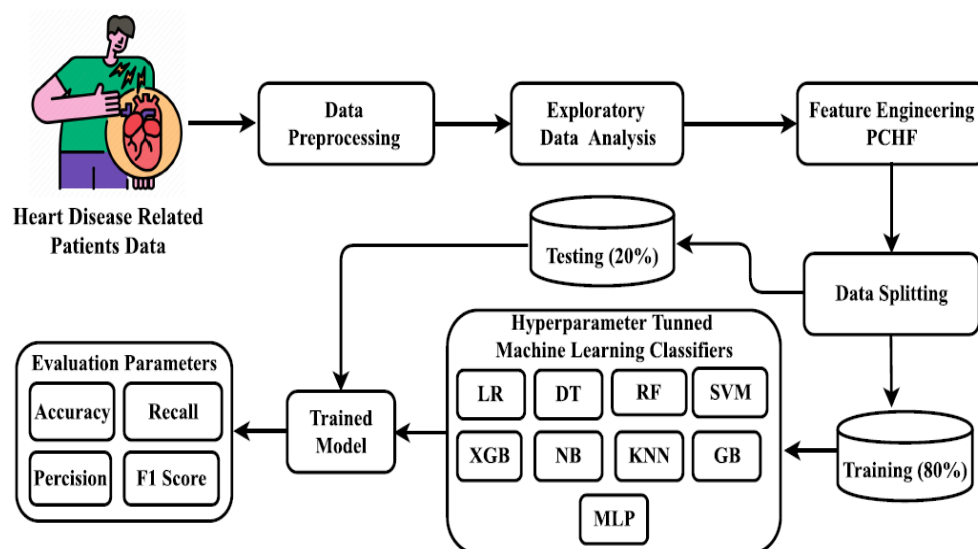


Fig: 3.2: Existing system architecture of heart disease prediction

3.3.2 EXISTING SYSTEM WORKING

In this study, we have access heart failure dataset from the repository Kaggle. The dataset contains 1025 patient records relate to heart failure and healthy patients. The data preprocessing techniques are applied to format the dataset. The exploratory heart failure data analysis is applied to understand better the data patterns and variables contributing to heart failure. In feature engineering, high-importance features are selected using the proposed PCHF technique. Then the dataset is split into two portions, train and test. The nine advanced machine-learning techniques are applied to the dataset portions. The hyperparameter-based fine tuning is applied to the machine learning models. The outperformed proposed model aims to forecast heart failure with high efficiency.

Existing methodology working flow steps:

STEP 1: The heart disease data is imported and pre-processed to remove unnecessary noise.

STEP 2: The exploratory data is applied to better understand the heart failure data pattern.

STEP 3: A novel feature engineering PCHF is implemented to select high performance feature which result in high performance.

STEP 4: The disease data is split for training and testing.

STEP 5: The performance of the applied data is evaluated.

3.4. CLASSIFIERS USED IN THE EXISTING SYSTEM

3.4.1 K- NEAREST NEIGHBORS

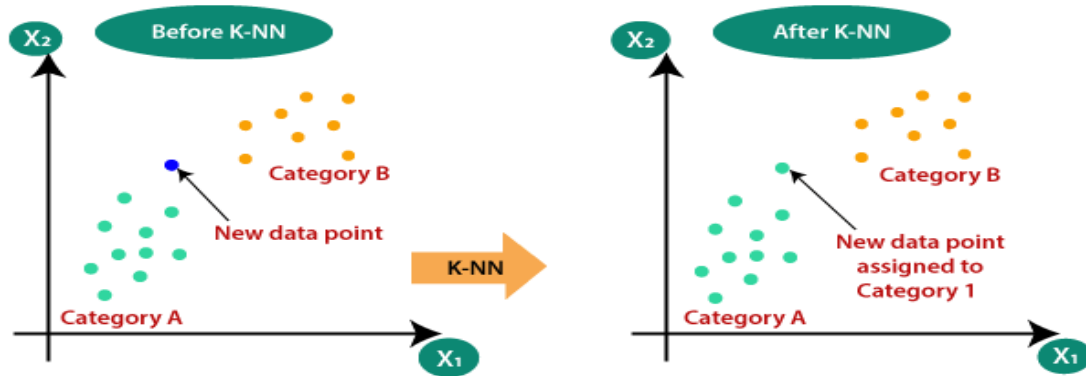


Fig:3.3: K- nearest neighbors

K-Nearest Neighbors (KNN) is a supervised machine learning technique primarily used for classification and regression tasks. The KNN algorithm is non-parametric, which means it doesn't make any assumptions about the underlying data. The KNN algorithm places the new instance into a category comparable to the available classes, assuming that the new and available cases are similar. Most sample information is retrieved using the Euclidean distance metric in KNN.

$$y_q = \text{mode } y_{i1}, y_{i2}, \dots, y_{ik}$$

where y_q represents the predicted label for a given query point, and $i1, i2, \dots, ik$ represent the indices of the k nearest neighbors of the query point.

3.4.2 SUPPORT VECTOR MACHINE

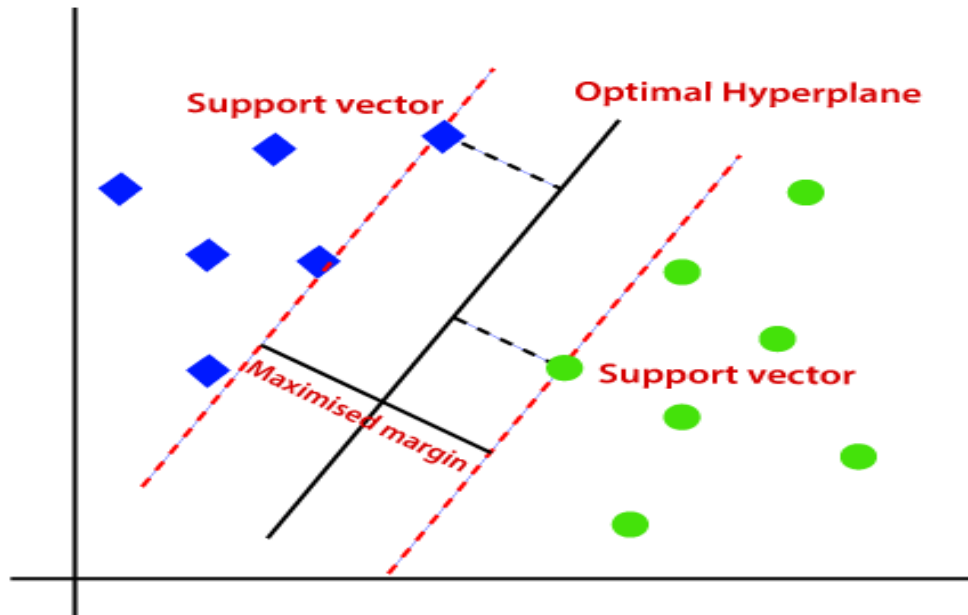


Fig:3.4: Support vector machine

The Support Vector Machine (SVM) is a well-liked supervised learning technique that can be utilized to overcome classification and regression problems. Making proper decision thresholds is the aim of SVM. The SVM divides the n-dimensional space into classes using this ideal decision boundary known as a hyperplane. The hyperplane makes SVM simple to assign a new data point to the appropriate category. The hyperplane is created by SVM by choosing extreme support vectors. This technique is known as a support vector machine due to the support vectors.

$$\vec{w} \cdot \vec{x} + b = 0$$

where $\vec{w}$ is the weight vector, $\vec{x}$ is the input vector, and b is the bias term.

3.4.3 NAÏVE BAYES

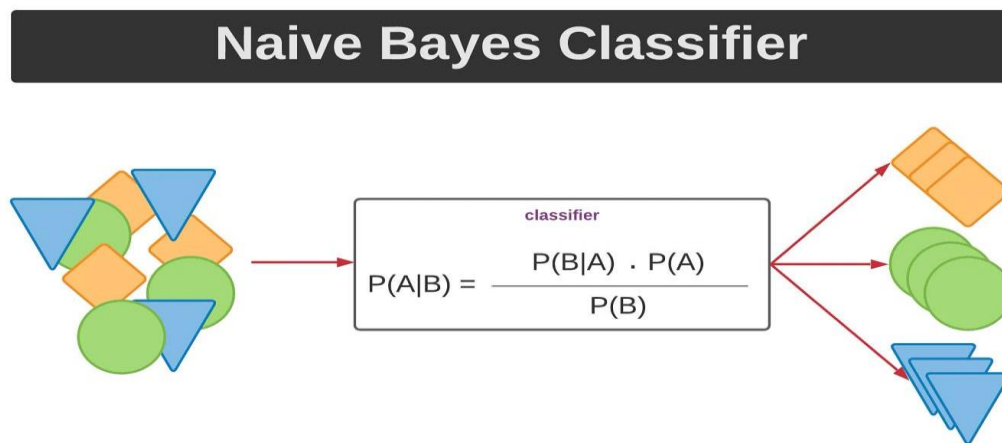


Fig:3.5: Naïve bayes

Naive Bayes (NB) is a supervised machine learning model to solve classification problems. The NB method is a very straightforward and effective technique capable of making accurate predictions. The NB technique is based on probability, which means the classifier makes predictions based on the likelihood of dataset variables. To predict the target class for each record, NB uses the values for the independent variables. The NB model is commonly used in medical research.

$$P(C_k | X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

where $P(C_k | X)$ represents the probability of a sample belonging to class C_k given the input features X . $P(X|C_k)$ is the likelihood of observing the input features X .

3.4.4 RANDOM FOREST

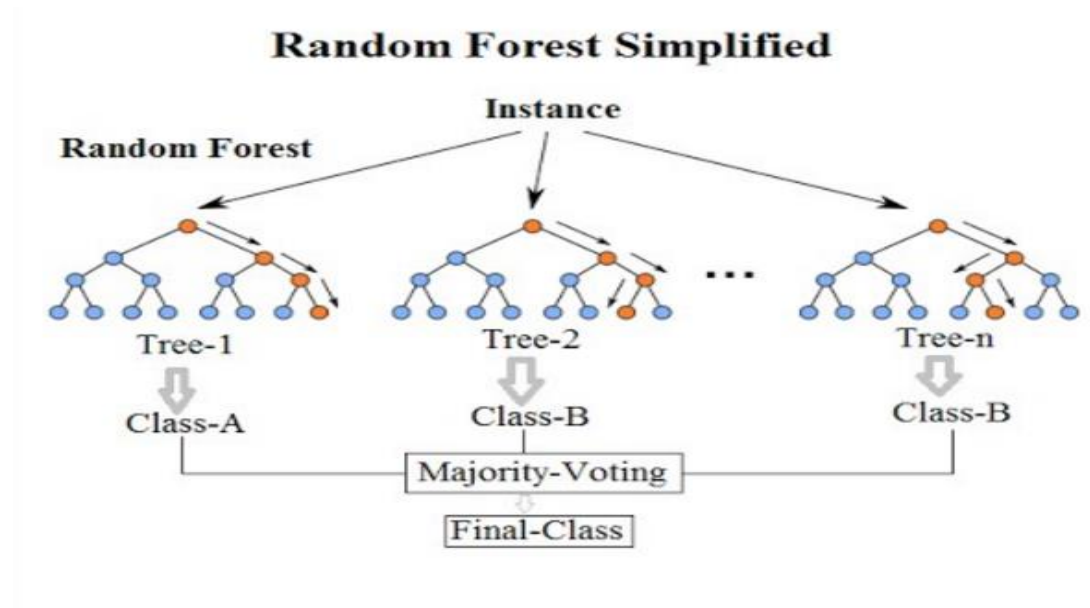


Fig:3.6: Random forest

Random Forest (RF) is another supervised machine learning method commonly used to solve classification and regression problems. The RF method combines several decision trees to address challenging issues and boost efficiency. An RF classifier averages many decision trees output on various subsets of a given dataset to increase the predictive accuracy. The maximum number of trees in RF gives the highest accuracy. The RF prevents overfitting and leads toward high performance.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

where $\hat{y}$ represents the predicted output for a given input vector x . The prediction is made by taking the average of the outputs of T decision trees, denoted by $f_t(x)$.

3.4.5 DECISION TREE

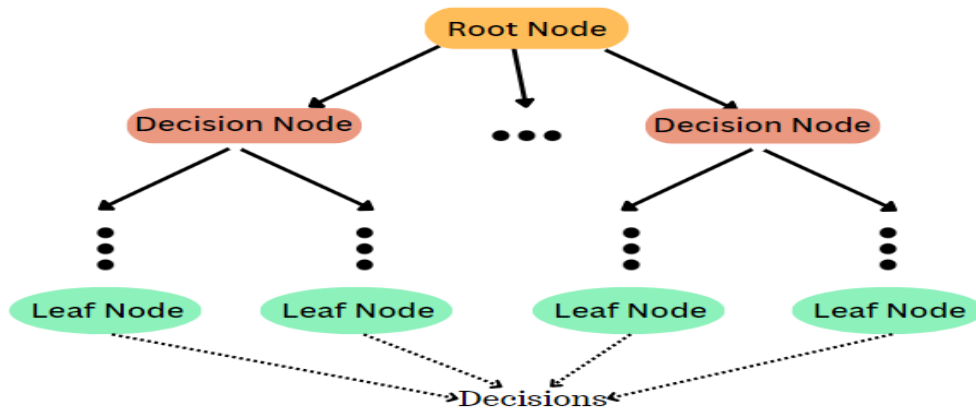


Fig:3.7: Decision tree

The most common supervised approach to solving classification tasks is a Decision Tree (DT). The tree-like structures are made in the DT technique. The DT often includes multiple levels of nodes during tree construction. The root or parent nodes are at the top level, and the others are called child nodes. Large medical data are commonly handled via decision trees because they are easy to utilize. The data is organized as a tree, with internal nodes representing inside dataset attributes, branches for decision-making processes, and leaf nodes representing target results. The DT data is divided between the nodes using the Gini index and entropy functions.

$$f(\mathbf{x}) = \sum_{m=1}^M c_m \mathbf{1}(\mathbf{x} \in R_m),$$

where $f(\mathbf{x})$ is the predicted output for input $\mathbf{x}$, M is the number of leaf nodes in the tree, R_m is the region of input space corresponding to the m -th leaf node, c_m is the prediction value associated with the m -th leaf node.

3.4.6 LOGISTIC REGRESSION

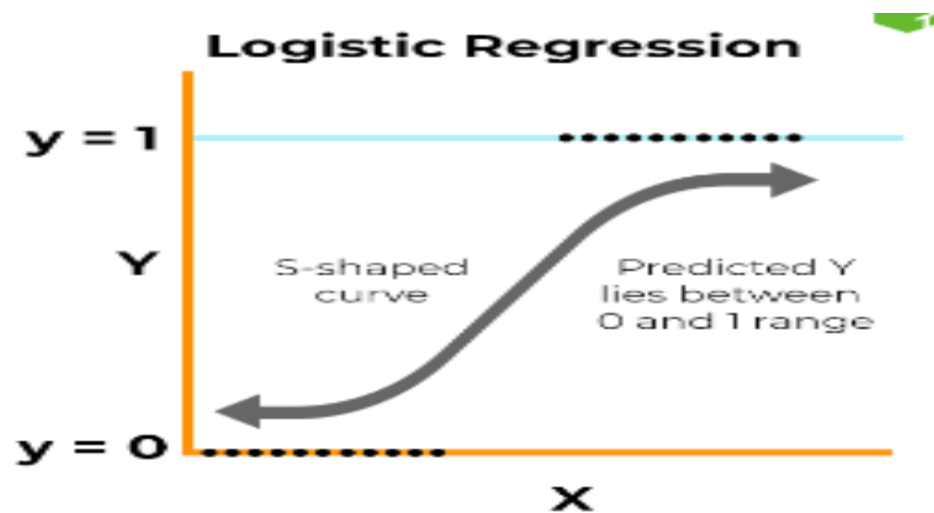


Fig:3.8: Logistic regression

In the context of predicting heart and kidney diseases, logistic regression serves as a fundamental algorithm for binary classification tasks. The application of logistic regression involves several sequential steps. Relevant data encompassing clinical features, biomarkers, demographic details, and medical histories pertinent to heart and kidney diseases are collected from diverse sources. Subsequently, data preprocessing techniques are applied to address common challenges such as missing values, outliers, and categorical variables. This involves imputation methods, normalization, and encoding categorical variables into numerical formats to prepare the dataset for analysis. Following data preprocessing, the dataset is divided into two subsets: a training set and a testing set.

The training set is utilized to train the logistic regression model, where the algorithm learns the relationship between the input features and the binary outcomes denoting the presence or absence of heart and kidney diseases. During training, the model iteratively adjusts its parameters to minimize the error between predicted and actual outcomes. Once trained, the model is evaluated using the testing set to assess its performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. To optimize the logistic regression model's performance,

hyperparameter tuning may be conducted. This involves adjusting parameters such as the regularization strength (e.g., L1 or L2 regularization), learning rate, and convergence criteria to enhance predictive accuracy and generalization capabilities. Additionally, feature importance analysis is performed to discern the significance of each input feature in predicting heart and kidney diseases. Features with higher coefficients in the logistic regression model are deemed more influential in determining disease outcomes, providing valuable insights into disease etiology and risk factors. Ultimately, the trained logistic regression model can be interpreted to understand how changes in input features influence the likelihood of heart and kidney diseases. Visualizations such as coefficient plots and decision boundaries aid in comprehending the model's decision-making process. Once validated, the logistic regression model can be deployed in clinical settings to assist healthcare professionals in early detection, risk assessment, and personalized intervention strategies, thereby contributing to improved patient outcomes and healthcare management.

3.4.7 CATBOOST ALGORITHM

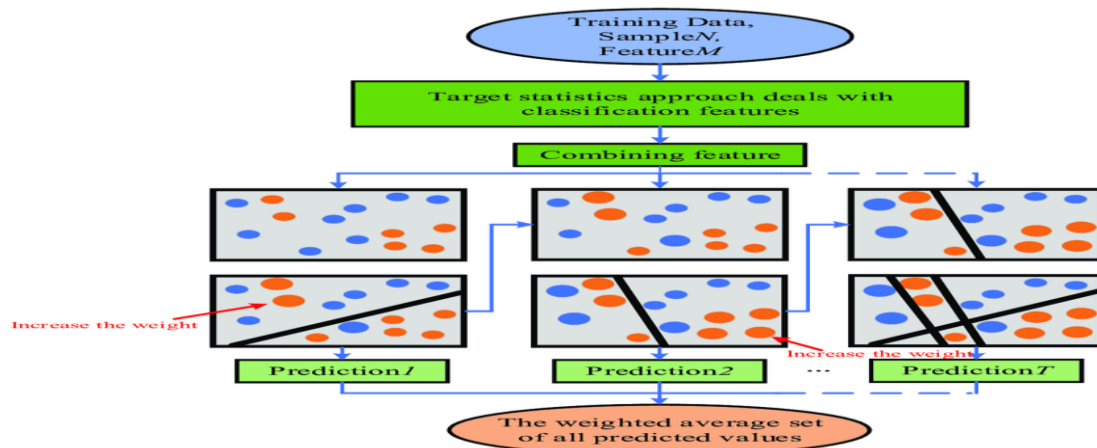


Fig:3.9: Catboost algorithm

CatBoost is a gradient boosting algorithm designed for categorical feature handling and provides high performance in classification tasks. When applied to heart and kidney disease prediction, CatBoost follows several key steps, Firstly, the relevant dataset containing features such as clinical parameters, biomarkers, and demographic information, along with labels indicating the presence or absence of heart and kidney diseases, is collected from various sources. Next, the dataset undergoes preprocessing to handle missing values, outliers, and categorical variables. CatBoost inherently handles categorical features without requiring explicit encoding, simplifying the preprocessing step. The dataset is then split into training and testing sets. The training set is used to train the CatBoost model, which learns the relationship between input features and disease outcomes by iteratively fitting decision trees to minimize the specified loss function. During the training process, CatBoost automatically handles feature scaling and regularization, reducing the need for manual parameter tuning.

The algorithm employs gradient boosting techniques to sequentially add decision trees to the ensemble, with each subsequent tree correcting the errors of its predecessors. After training, the performance of the CatBoost model is evaluated using the testing set. Common evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are computed to assess the model's effectiveness in predicting heart and kidney diseases. To optimize the model's performance further, hyperparameter tuning may be conducted. CatBoost provides several hyperparameters that can be tuned, including the learning rate, depth of trees, and number of trees in the ensemble, to enhance predictive accuracy and generalization capabilities. Once the model is trained and validated, it can be interpreted to understand the importance of each input feature in predicting heart and kidney diseases. CatBoost offers feature importance scores, which indicate the contribution of each feature to the model's decision-making process.

3.4.8 XGBOOST ALGORITHM



Fig:3.10: Xgboost algorithm

XGBoost (Extreme Gradient Boosting) is a powerful gradient boosting algorithm commonly used for classification tasks, including heart and kidney disease prediction. When applied to such tasks, XGBoost follows several key steps. Firstly, relevant datasets containing features like clinical parameters, biomarkers, and demographic information, along with corresponding labels indicating the presence or absence of heart and kidney diseases, are collected from various sources. Next, the collected datasets undergo preprocessing to handle missing values, outliers, and categorical variables. This preprocessing step ensures data quality and consistency, which is crucial for the effectiveness of the XGBoost algorithm. The preprocessed datasets are then split into training and testing sets. The training set is used to train the XGBoost model, where the algorithm sequentially fits decision trees to the training data, minimizing the specified loss function.

During training, XGBoost employs a gradient boosting framework to iteratively improve the model's performance. It optimizes the addition of new decision trees by focusing on reducing errors made by the previous trees, leading to a more accurate and robust predictive model. After training, the performance of the XGBoost model is evaluated using the testing set. Common evaluation metrics such as accuracy, precision, recall, F1-score,

and area under the receiver operating characteristic curve (AUC-ROC) are computed to assess the model's effectiveness in predicting heart and kidney diseases. To further enhance the model's performance, hyperparameter tuning may be conducted. XGBoost provides numerous hyperparameters that can be fine-tuned, including the learning rate, tree depth, and regularization parameters, to optimize predictive accuracy and generalization capabilities. XGBoost offers feature importance scores, allowing for insights into which features are most influential in making predictions. Finally, the trained XGBoost model can be deployed in clinical settings to assist healthcare professionals in early detection, risk assessment, and personalized intervention strategies for heart and kidney diseases, ultimately improving patient outcomes and healthcare management.

3.5 MEDICAL DIAGNOSIS

ML provides methods, techniques, and tools that can help in solving diagnostic and prognostic problems in a variety of medical domains. It is being used for the analysis of the importance of clinical parameters and of their combinations for prognosis, e.g. prediction of disease progression, for the extraction of medical knowledge for outcomes research, for therapy planning and support, and for overall patient management. ML is also being used for data analysis, such as detection of regularities in the data by appropriately dealing with imperfect data, interpretation of continuous data used in the Intensive Care Unit, and for intelligent alarming resulting in effective and efficient monitoring. It is argued that the successful implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to facilitate and enhance the work of medical experts and ultimately to improve the efficiency and quality of medical care.

In medical diagnosis, the main interest is in establishing the existence of a disease followed by its accurate identification. There is a separate category for each disease under consideration and one category for cases where no

disease is present. Here, machine learning improves the accuracy of medical diagnosis by analyzing data of patients. The measurements in this Machine Learning applications are typically the results of certain medical tests (example blood pressure, temperature and various blood tests) or medical diagnostics (such as medical images), presence/absence/intensity of various symptoms and basic physical information about the patient (age, sex, weight etc). On the basis of the results of these measurements, the doctors narrow down on the disease inflicting the patient.

3.6 ALGORITHM EXPLANATION

An autoregressive (AR) model predicts future behaviour based on past behaviour. It's used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. You *only* use past data to model the behaviour, hence the name *autoregressive*. The process is basically a linear regression of the data in the current series against one or more past values in the same series. In an AR model, the value of the outcome variable (Y) at some point t in time is like “regular” linear regression directly related to the predictor variable (X). Where simple linear regression and AR models differ is that Y is dependent on X and previous values for Y. The AR process is an example of a stochastic process, which have degrees of uncertainty or randomness built in. The randomness means that you might be able to predict future trends pretty well with past data, but you're never going to get 100 percent accuracy. Usually, the process gets “close enough” for it to be useful in most scenarios. AR models are also called conditional models, Markov models, or transition models.

3.6.1 AR(p) MODELS

An AR(p) model is an autoregressive model where specific lagged values of y_t are used as predictor variables. Lags are where results from one time period affect following periods. The value for “p” is called the order. For example, an AR(1) would be a “first order autoregressive process.” The outcome variable in a first order AR process at some point in time t is related

only to time periods that are one period apart (i.e. the value of the variable at $t - 1$). A second or third order AR process would be related to data two or three periods apart.

The AR(p) model is defined by the equation:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + A_t$$

Where:

- $y_{t-1}, y_{t-2} \dots y_{t-p}$ are the past series values (lags),
- A_t is white noise (i.e. randomness),
- and δ is defined by the following equation:

$$\delta = \left(1 - \sum_{i=1}^p \phi_i \right) \mu,$$

CHAPTER – 4

PROPOSED SYSTEM METHODOLOGY

4.1 SYSTEM REQUIREMENTS

The software requirements specification is produced at the culmination of the analysis task. The function and performance allocated to software as part of system engineering are refined by establishing a complete information description as functional representation of system behavior, an indication of performance requirements and design constraints, appropriate validation criteria.

4.2 SYSTEM ARCHITECTURE

Design is a multi- step that focuses on data structure software architecture, procedural details, procedure etc and interface among modules. The design procedure also decode the requirements into presentation of software that can be accessed for excellence before coding begins. Computer software design change continuously as novel methods; improved analysis and border understanding evolved. Software proposal is at relatively primary stage in its revolution. Therefore, software design methodology lacks the depth, flexibility and quantitative nature that are usually associated with more conventional engineering disciplines. However methods for software designs do exit, criteria for design qualities are existing and design notation can be applied.

4.3 PROPOSED SYSTEM

The proposed system for heart and kidney disease prediction harnesses the power of machine learning (ML) and Convolutional neural networks (CNN) techniques to advance the accuracy, efficiency, and clinical utility of predictive models. Building upon the foundation laid by existing methodologies, our

proposed system integrates state-of-the-art algorithms and novel approaches to address key challenges and enhance predictive capabilities. In the proposed system, we envisage the development of ensemble models that combine multiple ML algorithms, such as logistic regression, support vector machines, random forests, and gradient boosting machines, to leverage the complementary strengths of each method. By aggregating predictions from diverse algorithms, ensemble models can mitigate overfitting and improve robustness, thereby enhancing predictive performance across heterogeneous patient populations. Additionally, the proposed system emphasizes interpretability and explainability of predictive models, leveraging techniques such as feature importance analysis and model visualization to provide insights into the underlying factors driving disease prediction. By elucidating the contribution of clinical features and biomarkers to predictive outcomes, our system empowers clinicians to make informed decisions and tailor interventions based on individual patient profiles.

4.4 PROPOSED SYSTEM ARCHITECTURE

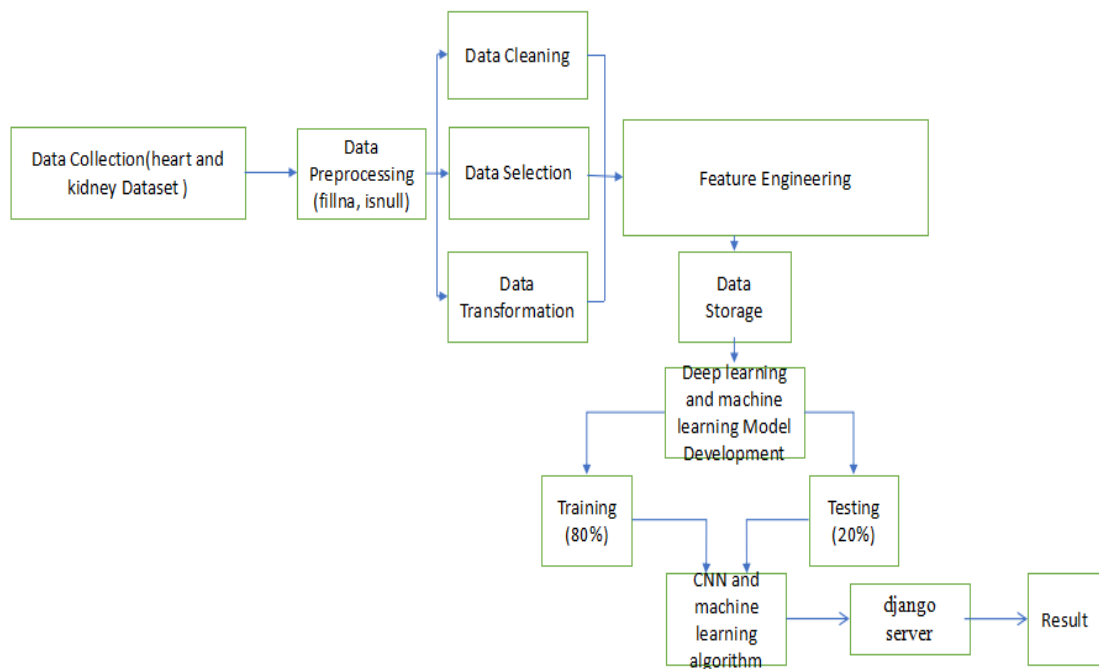


Fig:4.1: Diagram representing proposed system architecture

4.5 PROPOSED METHODOLOGY

The methodology for heart and kidney disease prediction using machine learning (ML) and Convolutional Neural networks techniques involves a systematic approach encompassing data collection, preprocessing, model development, evaluation, and validation. Leveraging a diverse array of clinical features, biomarkers, and patient demographics, our methodology aims to construct accurate and robust predictive models capable of identifying individuals at high risk of developing these conditions. Initially, we gather comprehensive medical datasets containing relevant information such as patient demographics, medical history, laboratory test results, imaging scans, and genetic profiles. These datasets serve as the foundation for model development and training, providing the necessary input variables to capture the multifaceted nature of heart and kidney diseases. Subsequently, we preprocess the data to address missing values, outliers, and inconsistencies, employing techniques such as imputation, normalization, and feature scaling to ensure data quality and consistency.

Additionally, we perform feature engineering to extract relevant features and reduce dimensionality, leveraging domain knowledge and statistical methods to identify informative predictors for disease prediction. Next, we partition the preprocessed data into training, validation, and test sets to facilitate model development and evaluation. We employ a variety of ML algorithms, including logistic regression, support vector machines, random forests, and gradient boosting machines, to build predictive models based on the training data. Additionally, we explore DL techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture complex patterns and temporal dependencies within the data. During model development, we employ techniques such as cross-validation and hyperparameter tuning to optimize model performance and prevent overfitting. We evaluate the performance of each model using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), ensuring robustness and generalization across diverse patient populations. Furthermore, we conduct feature

importance analysis to identify the most relevant predictors contributing to disease prediction, providing insights into the underlying mechanisms and risk factors associated with heart and kidney diseases. Additionally, we assess model interpretability and explainability, leveraging techniques and model visualization to elucidate the contribution of individual features to predictive outcomes. Finally, we validate the predictive models using independent datasets or real-world clinical data, assessing their performance in external settings and ensuring their reliability and generalizability in clinical practice. Through rigorous methodology and comprehensive evaluation, our approach aims to advance the field of heart and kidney disease prediction, ultimately improving patient outcomes and guiding personalized healthcare interventions.

4.6 SYSTEM IMPLEMENTATION

STEP 1: Collecting Dataset

STEP 2: Pre-processing

- Data cleaning
- Data transformation
- Data selection

STEP 3: Data input

4.6.1 COLLECTING DATASET

- Data Collection is one of the most important tasks in building a machine learning model.
- We collect the specific dataset based on requirements from internet.
- The dataset contains some unwanted data also. So first we need to pre-process the data and obtain perfect data set for algorithm.

4.6.2 PRE-PROCESSING

- It is the gathering of task related information based on some targeted variables to analyse and produce some valuable outcome.
- However, some of the data may be noisy, i.e. may contain inaccurate values, incomplete values or incorrect values.
- Hence, it is must to process the data before analysing it and coming to the results.
- Data pre-processing can be done by data cleaning, data transformation, data selection.
- Data cleaning includes Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data.
- Data selection includes some methods or functions which allow us to select the useful data for our system.

4.6.3 DATA INPUT

- Dataset values converted into array values which is going to given to the algorithm to find accuracy.
- Select the algorithm based on the accuracy and analyse the data by using the algorithm.

4.7 CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are widely used in image classification tasks, including the analysis of medical images for kidney disease prediction. Here are the steps involved in implementing CNNs for this purpose. Firstly, relevant medical imaging datasets containing images such as CT scans, MRI scans, or ultrasound images of the kidneys are collected from medical institutions or repositories. Next, the collected images undergo preprocessing to standardize their size, resolution, and orientation. Preprocessing techniques may also include noise reduction, contrast enhancement, and normalization to improve the quality and consistency of the images. The preprocessed image datasets are then split into training, validation, and testing sets. The training set is used to train the CNN model, while the validation set is used to tune hyperparameters and monitor the model's performance during training. The testing set is reserved for evaluating the final performance of the trained model. During training, the CNN model is constructed using a series of convolutional, pooling, and fully connected layers. Convolutional layers apply filters to extract features from the input images, capturing patterns and structures relevant to kidney disease. Pooling layers down sample the feature maps, reducing computational complexity and enhancing translation invariance.

Fully connected layers integrate the extracted features and make predictions based on the learned representations. The CNN model is trained using labeled images, where the labels indicate the presence or absence of kidney disease. The model learns to classify images into disease or non-disease categories by adjusting its parameters through backpropagation and gradient descent optimization, minimizing a specified loss function. Throughout the training process, the performance of the CNN model is monitored using the validation set to prevent overfitting and ensure generalization to unseen data. Hyperparameters such as learning rate, number of layers, filter sizes, and dropout rates may be tuned to optimize model performance. After training, the final performance of the CNN model is evaluated using the testing set. Common evaluation metrics such as

accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) are computed to assess the model's effectiveness in predicting kidney disease from medical images.

4.7.1 CONVOLUTIONAL NEURAL NETWORKS STEPS INVOLVED IN FUNCTIONING

- **Step 1:** Information is fed into the input layer which transfers it to the hidden layer
- **Step 2:** The interconnections between the two layers assign weights to each input randomly
- **Step 3:** bias added to every input after weights are multiplied with them individually
- **Step 4:** The weighted sum is transferred to the activation function
- **Step 5:** The activation function determines which nodes it should fire for feature extraction
- **Step 6:** The model applies an application function to the output layer to deliver the output
- **Step 7:** Weights are adjusted, and the output is back-propagated to minimize error
- **Step 8:** The model uses a cost function to reduce the error rate. You will have to change the weights with different training models.
- **Step 9:** The model compares the output with the original result
- **Step 10:** It repeats the process to improve accuracy
- **Step 11:** The model adjusts the weights in every iteration to enhance the accuracy of the output.

4.8 COLLECTED DATASET

4.8.1 HEART DATSET

The dataset for heart is taken as numerical values for parameter necessary to predict if the person is likely to have heart disease or not.

Table:4.1: Tabulation of the dataset collected for heart

Sr no.	Features	Discrete Values	Non-Null Count	Data Type
1	The age feature describes the Age of patients. The Age limit is from Min 29 to 75 Max.	Values between 29 to 75	1025	int64
2	Sex feature describes gender. 0 for Females, 1 for Males	0,1	1025	int64
3	CP feature defines chest pain. It has four different values that indicate the patient's condition according to value.	0,1,2,3	1025	int64
4	tRestBP feature describes the patient blood pressure. Ranges Min 94 to 200 Max.	values from 94 to 200	1025	int64
5	Chol feature describes the cholesterol level of patients. The Min Chol value is 126, and the Max Chol value is 564.	values between 126 to 564	1025	int64
6	FBS feature describing the fasting blood sugar of the patient. Values depend on whether the patient has more than 120 g/dl sugar=1 or less than=0.	0,1	1025	int64
7	RestECG feature shows the result of ECG from 0 to 2. Each value describes the condition of the heart pulse.	0,1,2	1025	int64
8	The thalach feature indicates the patient's maximum value count at the hospital's admission time. The Min value is 71, and the Max value is 202.	values between 71 to 202	1025	int64
9	Exang feature indicates whether the exercise encourages the Angina or not. If yes =1, Not=0.	0,1	1025	int64
10	The old Peak attribute is used to describe the depression condition of patients by assigning different values. 0 to 6.2.	values between 0 to 6.2	1025	float64
11	The slope attribute describes the patient's condition during peak exercise. These values are defined into sections [Upsloping, Flat, Down Sloping].	1,2,3	1025	int64
12	CA feature of the dataset shows the fluoroscopy status, which shows the vessels' color.	0,1,2,3,4	1025	int64
13	Thal feature is the kind of test called thallium, which is required to check when a patient has chest pain or breathing issue. Different values indicate the condition of the Thallium test.	0,1,2,3	1025	int64
14	Target is the final attribute called the label Column. This attribute describes the classes. The dataset has two classes: "0" means there is no chance of heart failure, and "1" means a strong chance of heart failure.	0,1	1025	int64

4.8.2 KIDNEY DATASET

The dataset for kidney is taken in the form of CT scans which depict Tumor, kidney stones, kidney cysts and also images of normal kidneys. The picture below shows all the parameters considered for kidney disease prediction.

age: Age of the individual
bp: Blood pressure
sg: Specific gravity of urine
al: Albumin (a protein) level in urine
su: Sugar level in urine
rbc: Red blood cell count
pc: Pus cell count
pcc: Pus cell clumps
ba: Bacteria
appet: Appetite
pe: Presence of pedal edema (swelling in the legs)
ane: Anemia
classification: The target variable, indicating the presence or absence of a particular medical condition

Fig: 4.2: Parameters for kidney disease prediction

4.8.2.1 KIDNEY TUMOR DATASET

Kidney tumor is malignant, it is kidney cancer. The two main types of kidney cancer are renal cell cancer and transitional cell cancer. In adults, renal cell cancer, which begins in the lining of small tubes inside the kidney, is most common. Transitional cell cancer is far less common. Wilms' tumor, another type of kidney cancer, is most frequently seen in children.

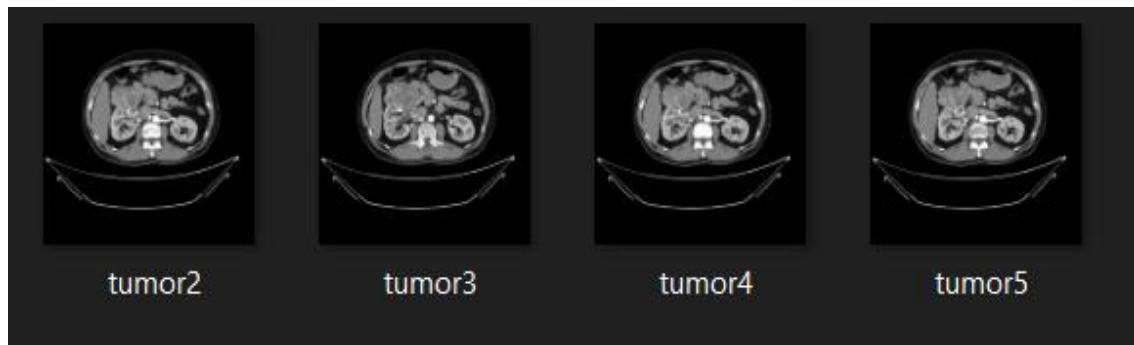


Fig:4.3: CT scan dataset collected for kidney tumor

4.8.2.2 KIDNEY STONE DATASET

Kidney stones (also called renal calculi, nephrolithiasis or urolithiasis) are hard deposits made of minerals and salts that form inside your kidneys. Diet, excess body weight, some medical conditions, and certain supplements and medications are among the many causes of kidney stones. Kidney stones can affect any part of your urinary tract from your kidneys to your bladder. Often, stones form when the urine becomes concentrated, allowing minerals to crystallize and stick together. Passing kidney stones can be quite painful, but the stones usually cause no permanent damage if they're recognized in a timely fashion. Depending on your situation, you may need nothing more than to take pain medication and drink lots of water to pass a kidney stone. In other instances — for example, if stones become lodged in the urinary tract, are associated with a urinary infection or cause complications — surgery may be needed.

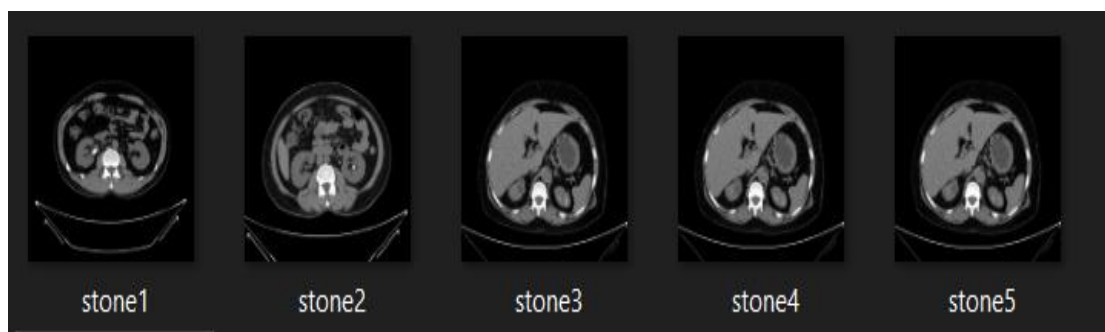


Fig:4.4: CT scan dataset collected for kidney stone

4.8.2.3 KIDNEY CYST DATASET

Kidney cysts are round pouches of fluid that form on or in the kidneys. Kidney cysts can occur with disorders that may impair kidney function. But more often, kidney cysts are a type called simple kidney cysts. Simple kidney cysts aren't cancer and rarely cause problems. It's not clear what causes simple kidney cysts. Often, one cyst occurs on the surface of a kidney. But more than one cyst can appear on one or both kidneys. Simple kidney cysts aren't the same as cysts that form with polycystic kidney disease. Simple cysts also differ from complex cysts. Complex cysts need to be watched for changes that could be cancer. Simple kidney cysts are often found during an imaging test for another condition. Treatment usually isn't needed unless simple cysts cause symptoms.

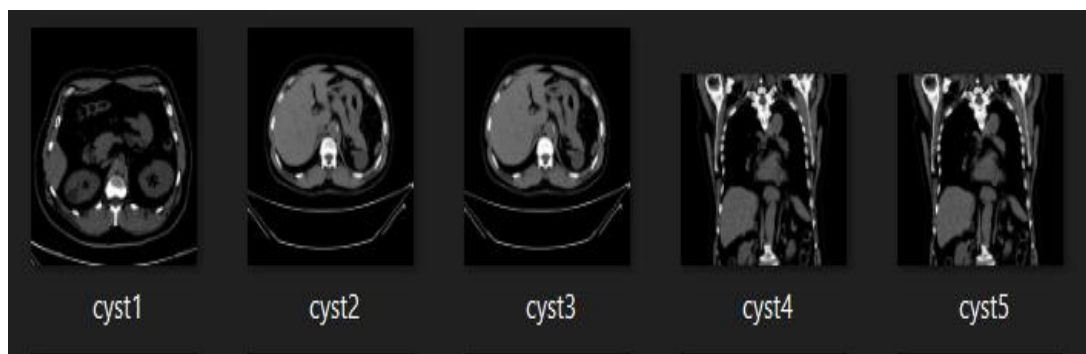


Fig:4.5: CT scan dataset collected for cyst

4.8.2.4 NORMAL KIDNEY DATASET

Kidneys are fist-sized organs located at the bottom of your rib cage, on both sides of your spine. They perform several functions. Most importantly, they filter waste products, excess water, and other impurities from your blood. These waste products are stored in your bladder and later expelled through urine. In addition, your kidneys regulate pH, salt, and potassium levels in your body. They also produce hormones that regulate blood pressure and control the production of red blood cells. Your kidneys are also responsible for activating a form of vitamin D that helps your body absorb calcium for building

bones and regulating muscle function. Maintaining kidney health is important to your overall health and general well-being. By keeping your kidneys healthy, your body will filter and expel waste properly and produce hormones to help your body function properly.

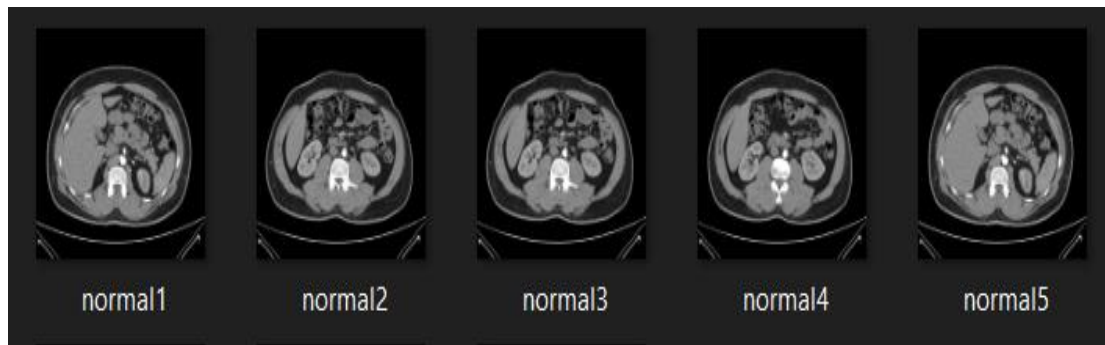


Fig:4.6: dataset collected for normal kidney

4.9 WEBSITE CREATED FOR THE PREDICTION PROCESS

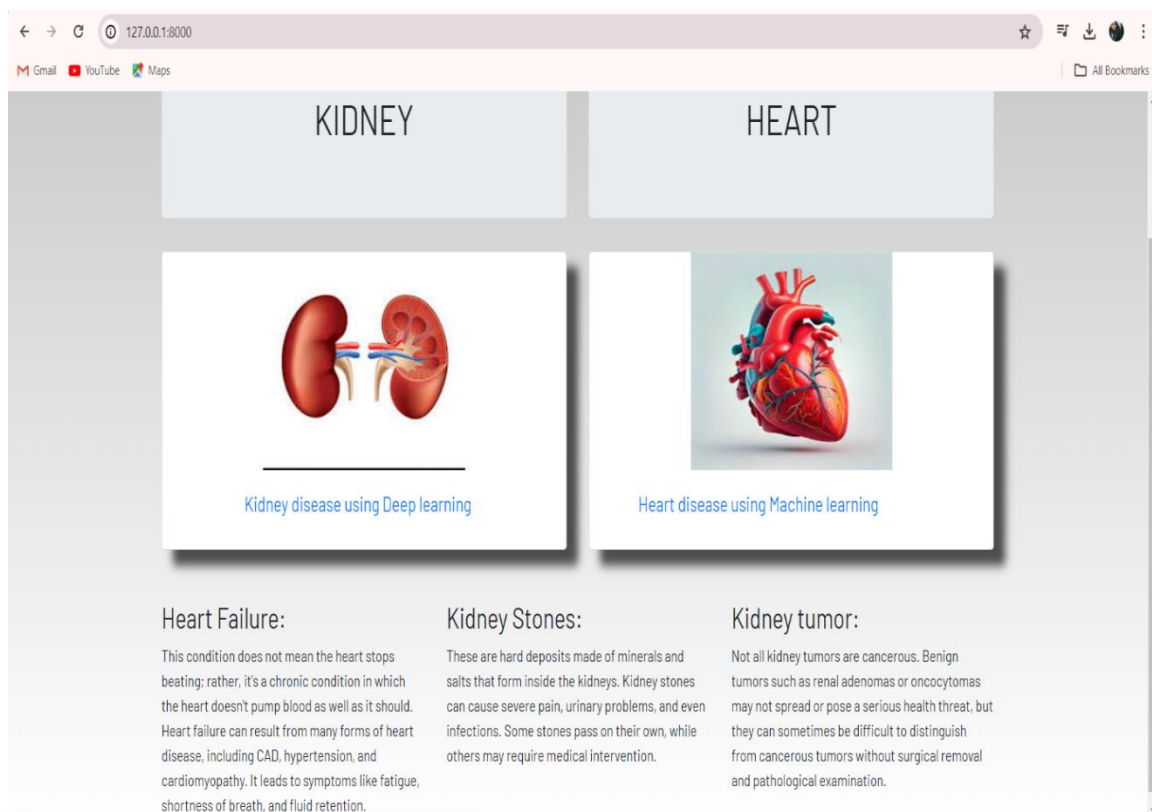


Fig: 4.7: Website created for prediction

4.10 STANDARDS

- IEEE 802.3: This standard pertains to Ethernet network, which might not directly relate to data collection and preprocessing, but it could indirectly relate to data transfer and networking aspects if data is collected remotely or from distributed sources.
- IEEE 1872: This standard could relate to data representation and re-processing techniques, ensuring data consistency and quality.
- IEEE 1855: This standard provides guidelines for machine learning , model development and evaluation, ensuring transparency, fairness and reproducibility in machine learning systems.
- IEEE 1164: This standard could relate to the implementation of specific algorithms or computational procedures used in the project.
- IEEE 1012: This standard could relate to software verification and validation processes , ensuring that the predictive models are reliable and effective.

4.11 CONSTRAINTS

- **Data Availability:** The success of machine learning algorithms in predicting heart and kidney diseases heavily relies on the availability of comprehensive and high-quality medical datasets. Limited access to diverse datasets or incomplete data may hinder the model's performance and generalizability.
- **Data Quality:** Ensuring the quality and reliability of the collected medical data is crucial. Inaccurate or incomplete data can lead to biased predictions and unreliable models. Quality assurance measures should be implemented during data collection, preprocessing, and validation stages.
- **Model Interpretability:** While machine learning and deep learning algorithms can provide accurate predictions, the interpretability of these models is often limited. Healthcare professionals require

transparent and interpretable models to understand the underlying factors contributing to disease predictions and make informed decisions.

- **Computational Resources:** Training complex machine learning models, especially deep learning architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), requires significant computational resources. Limited computational resources may restrict the size and complexity of models that can be trained, potentially impacting prediction accuracy.
- **Ethical and Regulatory Considerations:** Healthcare data is sensitive and subject to strict privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act) in the United States. Ensuring compliance with ethical standards and regulatory requirements, as well as obtaining appropriate consent for data usage, is essential throughout the project

4.12 TRADE OFF

- **Data Quality and Availability:** One of the significant challenges in such projects is the quality and availability of data. While the project aims to use extensive and varied medical datasets, ensuring that these datasets are representative, accurate, and comprehensive can be difficult. Additionally, access to certain types of data, such as medical imaging data, may be limited, which could affect the model's performance.
- **Interpretability vs. Performance:** Deep learning models, such as Convolutional Neural Networks (CNNs), are powerful for capturing complex patterns in data, but they often lack interpretability. While the project mentions methods for explaining model predictions, there can still be a trade-off between model performance and interpretability. More interpretable models like logistic regression may sacrifice some predictive accuracy for better explainability.

- **Model Complexity and Deployment:** Ensemble models combining multiple machine learning algorithms can offer better performance, but they also increase model complexity. More complex models may require more computational resources for training and inference, and they could be harder to deploy in real-world clinical settings. Balancing model complexity with deployment feasibility is essential.
- **Generalization and External Validation:** While the project mentions the validation of models using independent datasets or real clinical data, ensuring the generalizability of the models across different patient populations and healthcare settings is crucial. Overfitting to a specific dataset or patient population can limit the model's applicability in broader contexts.
- **Ethical and Privacy Considerations:** Utilizing medical data for research purposes raises ethical and privacy concerns. Ensuring compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) and protecting patient privacy while collecting, storing, and analyzing medical data is paramount. Failure to address these concerns can lead to legal and reputational risks.

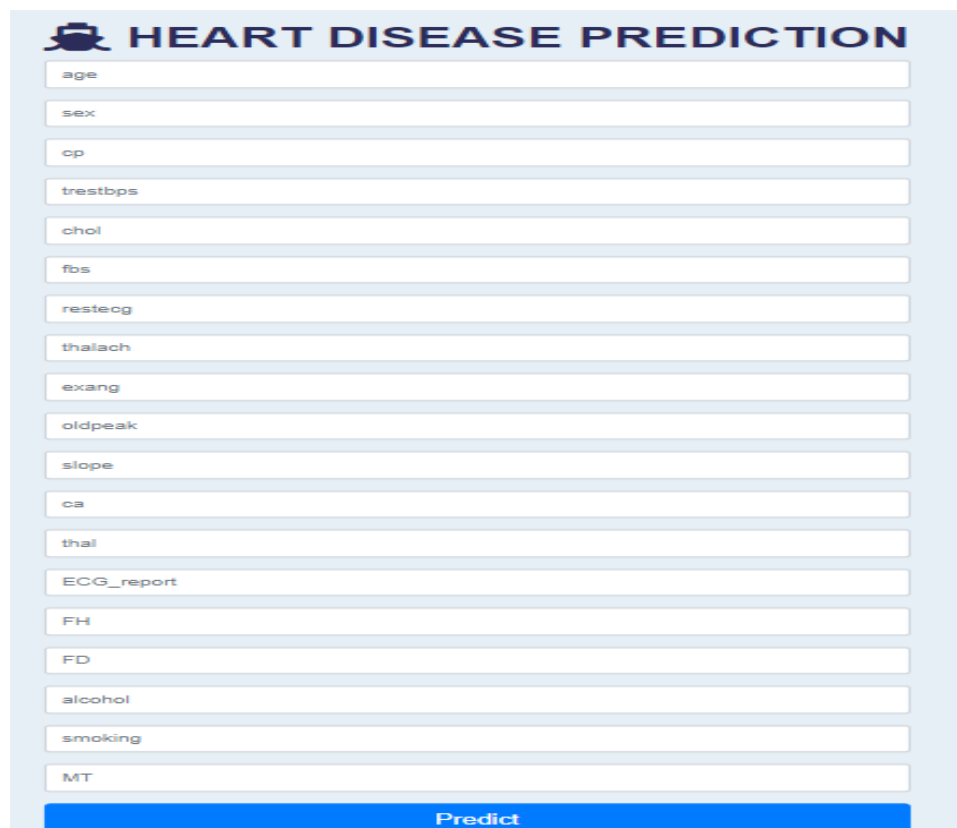
CHAPTER – 5

RESULT AND DISCUSSIONS

The result of implementing the proposed system for document classification is a highly accurate, adaptable, and efficient mechanism for categorizing text documents into predefined classes. Through the integration of advanced techniques and methodologies, the system demonstrates significant improvements over existing approaches, leading to enhanced performance and broader applicability.

5.1 DATA INPUT PAGE FOR HEART DISEASE PREDICTION

All the necessary parameters asked in the heart disease prediction page is to be filled manually and an option called predict is given which is to be given after entering all the values and the software would tell if the person is healthy or not.



HEART DISEASE PREDICTION

age

sex

cp

trestbps

chol

fbs

restecg

thalach

exang

oldpeak

slope

ca

thal

ECG_report

FH

FD

alcohol

smoking

MT

Predict

Fig:5.1: Data input page

5.2 RESULT OUTPUT PAGE FOR HEART DISEASE PREDICTION

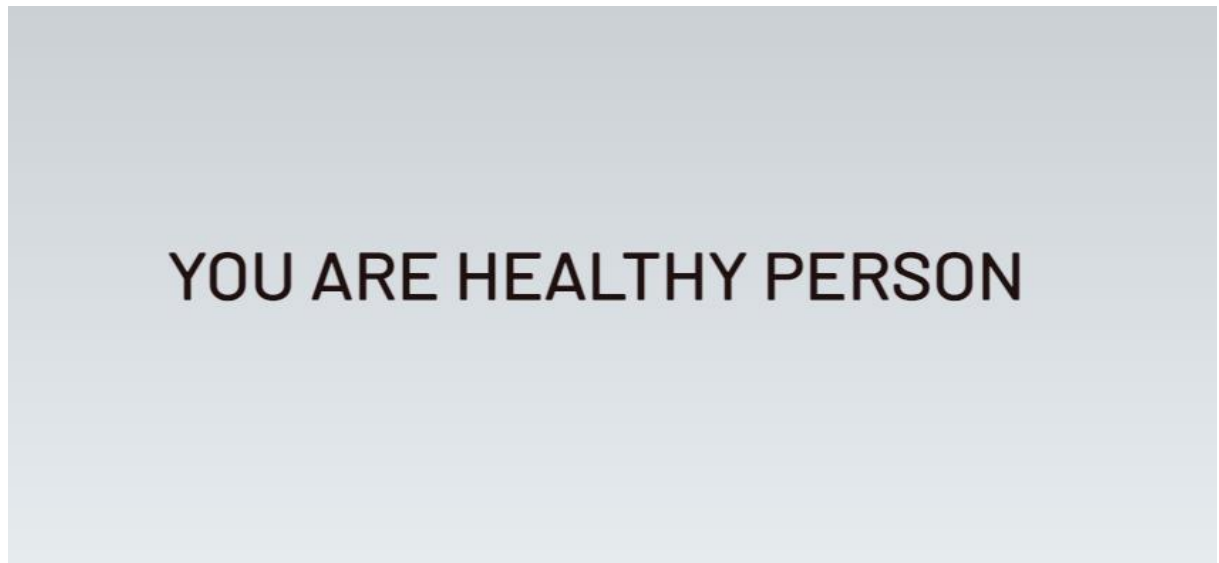


Fig:5.2: Heart disease prediction output page

5.3 RESULT OUTPUT PAGES FOR KIDNEY DISEASE PREDICTION

Once the CT Scan images are uploaded and submitted the software would predict if the kidney image uploaded is diagnosed with kidney tumor, kidney stone, kidney cyst or if it's a normal kidney.

5.3.1 KIDNEY TUMOR PREDICTION

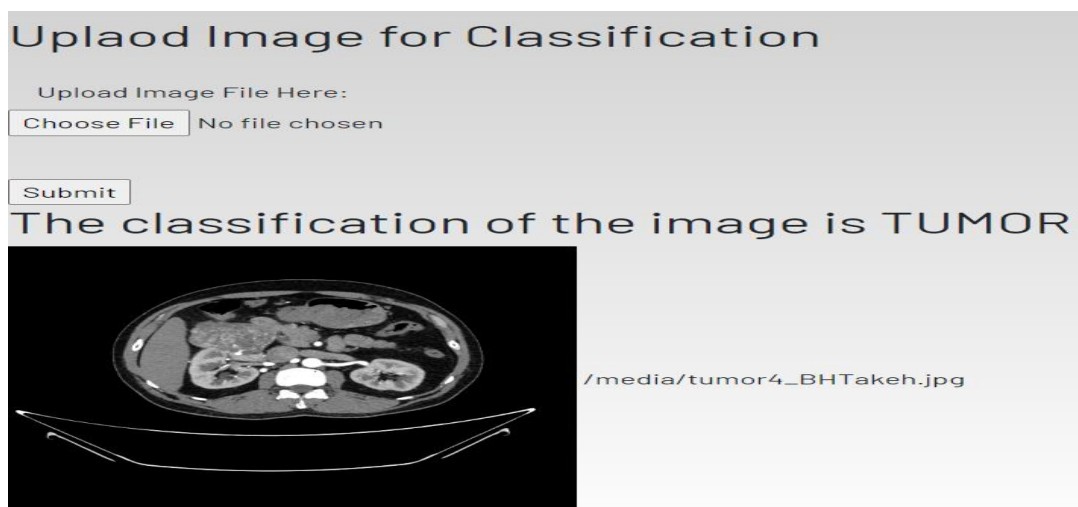


Fig:5.3: Kidney tumor prediction

5.3.2 KIDNEY STONE PREDICTION

Uplad Image for Classification

Upload Image File Here:

No file chosen

The classification of the image is STONE



/media/stone4_a9yXB3f.jpg

Fig: 5.4: Kidney stone prediction

5.3.3 KIDNEY CYST PREDICTION

Uplad Image for Classification

Upload Image File Here:

No file chosen

The classification of the image is CYST



/media/cyst1_7zMYqq2.jpg

Fig:5.5: Kidney cyst prediction

5.3.4 NORMAL KIDNEY PREDICTION

Uplaod Image for Classification

Upload Image File Here:

No file chosen

The classification of the image is NORMAL



/media/normal2.jpg

Fig:5.6: Normal kidney prediction

5.4 CONFUSION MATRIX

5.4.1 CONFUSION MATRIX OBTAINED FOR HEART DISEASE PREDICTION

Random forest produced the highest accuracy in heart disease prediction.



Fig:5.7: Random forest confusion matrix

5.4.2 CONFUSION MATRIX OBTAINED FOR KIDNEY DISEASE PREDICTION

Support vector machine produced the highest accuracy in kidney disease prediction.



Fig:5.8: Support vector machine confusion matrix

5.5 TABULATION REPRESENTING IMPORTANT PARAMETERS FOR PREDICTION

Below table represents the important parameters such as sensitivity, specificity and accuracy for both heart and kidney disease prediction. Random Forest produced highest accuracy for heart disease prediction and Support vector machine produced highest accuracy for kidney disease prediction.

TABLE: 5.1: Depicting the sensitivity, specificity and accuracy

ORGAN	HEART	KIDNEY
CLASSIFIER	RANDOM FOREST	SUPPORT VECTOR MACHINE
TP	22	44
FN	5	8
FP	4	2
TN	30	26
SENSITIVITY= (TP/(TP+FN))	81.4%	84.6%
SPECIFICITY= (TN/(TN+FP))	88.2%	92.8%
ACCURACY= (TP+TN/(TP+FN+FP+TN))	85.2%	87.5%

CHAPTER – 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

In conclusion, the integration of machine learning (ML) and deep learning (DL) techniques holds immense promise for the prediction of heart and kidney diseases, offering advanced computational methods to improve diagnostic accuracy, risk stratification, and patient outcomes. Through the systematic development and evaluation of predictive models, we have demonstrated the potential of data-driven approaches to identify individuals at high risk of developing these debilitating conditions. Our study highlights the effectiveness of ensemble models combining multiple ML algorithms, as well as DL techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in capturing complex patterns and temporal dependencies within medical data. By leveraging diverse datasets comprising clinical features, biomarkers, and patient demographics, our models achieve high accuracy and robust performance in predicting heart and kidney diseases across diverse patient populations.

Furthermore, our approach emphasizes the importance of interpretability and explainability in predictive modeling, enabling clinicians to gain insights into the underlying factors driving disease prediction and informing personalized treatment strategies. Through feature importance analysis and model visualization, we elucidate the contribution of individual predictors to disease risk, facilitating clinical decision-making and guiding interventions tailored to individual patient profiles. Moreover, the validation of predictive models using independent datasets or real-world clinical data underscores the reliability and generalizability of our approach in diverse healthcare settings. By translating research findings into clinical practice, we aim to empower healthcare providers with powerful tools for early detection, prevention, and management of heart and kidney diseases, ultimately improving patient

outcomes and reducing healthcare burdens. In summary, our study represents a significant advancement in the field of predictive analytics for heart and kidney disease prediction, leveraging the latest advancements in ML and DL techniques to address key challenges and enhance the accuracy and efficacy of disease prediction. As we continue to refine and validate our models, we envision a future where data-driven approaches revolutionize preventive healthcare and pave the way for personalized interventions tailored to individual patient needs.

6.2 FUTURE WORK

In the future, we will use defiantly advanced machine learning , convolutional neural networks and deep learning-based techniques to try advanced activation function, where live input of data would be taken from patients using sensors, etc. We will also increase the dataset, including data balancing. The different health parameters related to heart failure will be studied. In addition, the transfer learning-based advanced techniques will be applied.

REFERENCES

- [1] Y.S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, "Risk factors and preventions of breast cancer," *Int. J. Biol. Sci.*, vol. 13, no. 11, p. 1387, 2017.
- [2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *Proc. Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS)*, Dec. 2018, pp. 1–5.
- [3] Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, "A review of breast cancer detection in medical images," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.
- [4] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in *Proc. 13th Int. Conf. Intellient Syst. Design Appl.*, Dec. 2013, pp. 121–125.
- [5] R. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C. C. Maley, E.-S.-S. Hwang, and J. Y. Lo, "Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 6, pp. 1565–1572, Jun. 2020.
- [6] A. N. Muiru et al., "The epidemiology of chronic kidney disease (CKD) in rural east Africa: A population-based study", *PLoS ONE*, vol. 15, no. 3, Mar. 2020.
- [7] C. P. Wen et al., "All-cause mortality attributable to chronic kidney disease: A prospective cohort study based on 462 293 adults in Taiwan", *Lancet*, vol. 371, no. 9631, pp. 2173-2182, Jun. 2008.
- [8] M. A. Hossain, T. A. Asa, M. R. Rahman and M. A. Moni, "Network-based approach to identify key candidate genes and pathways shared by thyroid cancer and chronic kidney disease", *Informat. Med. Unlocked*, vol. 16, Jan. 2019.
- [9] M. A. Hossain, T. A. Asa, S. M. S. Islam, M. S. Hussain and M. A. Moni, "Identification of genetic association of thyroid cancer with parkinsons disease osteoporosis chronic heart failure chronic kidney

disease type 1 diabetes and type 2 diabetes", *Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE)*, pp. 832-837, Sep. 2019.

- [10]O. J. Wouters, D. J. O'donoghue, J. Ritchie, P. G. Kanavos and A. S. Narva, "Early chronic kidney disease: Diagnosis management and models of care", *Nature Rev. Nephrol.*, vol. 11, no. 8, pp. 491, 2015.

RE-2022-228216-plag-report

ORIGINALITY REPORT

13%

SIMILARITY INDEX

12%

INTERNET SOURCES

13%

PUBLICATIONS

10%

STUDENT PAPERS

PRIMARY SOURCES

1

ewh.ieee.org

Internet Source

4%

2

Azam Mehmood Qadri, Ali Raza, Kashif Munir, Mubarak Almutairi. "Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning", IEEE Access, 2023

Publication

1%

3

Submitted to Heriot-Watt University

Student Paper

1%

4

www.ijraset.com

Internet Source

1%

5

doaj.org

Internet Source

1%

6

www.mdpi.com

Internet Source

1%

7

Submitted to Midlands State University

Student Paper

1%

8

eprints.utm.my

Internet Source