NOVEL MACHINE LEARNING APPROACH FOR PREDICTING CHRONIC KIDNEY DISEASES

A Thesis submitted

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE AND ENGINEERING

By

Mr. SREEJI S
17SCSE301022

Supervisor

Dr. B. BALAMURUGAN
Professor

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY
Plot No 2, Sector 17-A Yamuna Expressway
Greater Noida, Uttar Pradesh
INDIA

OCTOBER, 2021

CANDIDATE DECLARATION

I hereby certify that the work which is being presented in the thesis, entitled "**NOVEL MACHINE LEARNING APPROACH FOR PREDICTING CHRONIC KIDNEY DISEASES**" in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Faculty of Computer Science and Engineering and submitted in Galgotias University, Uttar Pradesh is an authentic record of my own work carried out during a period from December 2017 under thesupervision of **Dr. B. BALAMURUGAN**, Professor, School of Computing Science and Engineering, Galgotias University.

The matter embodied in this thesis has not been submitted by me for the award of any other degree or from any other University/Institute.

(Sreeji. S)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Dr. B. BALAMURUGAN)
Supervisor
School of Computing Science and Engineering

Galgotias University Uttar Pradesh
School of Computing Science & Engineering

CERTIFICATE

This is to certify that **Mr. SREEJI.S** has presented his pre-submission seminar of the thesis entitled **" NOVEL MACHINE LEARNING APPROACH FOR PREDICTING CHRONIC KIDNEY DISEASES "** before the committee and summary is approved and forwarded to School Research Committee of School of Computing Science & Engineering, in the Faculty of Engineering & Technology, Galgotias University Uttar Pradesh.

Dean – SCSE **Dean – Ph.D & PG**

The Ph.D. Viva-Voice examination of **Sreeji. S** Research Scholar, has been held on ____

Supervisor **External Examiner**

APPROVAL SHEET

This thesis/dissertation/report entitled " **NOVEL MACHINE LEARNING APPROACH FOR PREDICTING CHRONIC KIDNEY DISEASES** **“** by Sreeji. S is approved for the degree of Doctor of Philosophy

Examiner **Supervisor** **Chairman**

STATEMENT OF THESIS PREPARATION

1. Thesis title: Novel Machine learning approach for predicting chronic kidney diseases
2. Degree for which the thesis is submitted: Doctor of Philosophy in CSE
3. The thesis preparation was done based on the thesis guide.
4. The thesis format and specifications are keenly followed while preparing the thesis.
5. The guidelines for the arrangement of the thesis is adhered carefully.
6. The thesis is prepared in such a way that it does not have any plagiarism from any sources.
7. All the references have been cited appropriately within the document.
8. The thesis is original and has no reference of being submitted anywhere for the award of the degree.

(Signature of the student)

Name: Sreeji. S
Roll No. 17SCSE301022

ACKNOWLEDGEMENT

Being an Assistant Professor and doing research for the degree of Ph. D in Galgotias University was quite magnificent and challenging experience for me. In all these years, many people directly or indirectly contributed in shaping up my career. It was hardly possible for me to complete my doctoral work without the precious and invaluable support of these personalities. I would like to give my small tribute to all those people.

Initially, I would like to record my deep gratitude to my supervisor, Dr. Balamurugan, for his valuable guidance, enthusiasm and overfriendly nature that helped me a lot to complete my research work.

I must owe a special debt of gratitude to Hon'ble Chancellor Mr.Suneel Galgotia, Mr.Dhruv Galgotia, CEO, and Honb'le Vice Chancellor, Galgotias University for their valuable cooperation.

I express my gratitude to Dr Munish Sabharwal, Dean of School of Computing Science and Engineering, Dr. Naresh kumar , Dean PG & Ph.d for his guidance and moral support.

I would like to convey my deep regard to my dad for his wise counsel and indispensable advice that always encouraged me to work hard for completion of the thesis. Finally, this work would not have been possible without the confidence, endurance and support of my family. My highest gratitude goes to my parent and in-laws for their relentless supports, blessing and encouragement. Special mention goes to my wife, Bemi Jerald, whose time I stole to write this thesis.

Mr. Sreeji. S

CONTENTS

List of Figures

List of Tables

List of Abbreviations

CKD	Chronic Kidney Disease
CVD	Cardio-Vascular Disease
CDSS	Clinical Decision Support System
CbH	Choice based Hierarchy
ANFIS	Adaptive Neuro-Fuzzy Inference System
GFR	Glomerular filtration rate
MDRD	Modification of diet in Renal Disease equation
CD	Chronic diseases
ML	Machine Learning
AI	Artificial Intelligence
RL	Reinforcement learning
EHR	Electronic health records
CNN	Computational Neural Network
IgAN	Iimmunoglobulin A nephropathy
DKDs	Diabetic kidney diseases
AKI	Acute kidney injury
MDRD	Modification of Diet in Renal Disease
DTPA	Diethylene triamine pentaacetic acid
ESRD	End Stage Renal Disease
FOAM	fuzzy optimal Associative Memory
FuRES	fuzzy rule-building expert system
IHFCM	Improved Hybrid Fuzzy C-Means
WAELI	Weighted Average Ensemble Learning Imputation
SVM	support vector machine
kNN	k-nearest neighbor
CFS	Correlation-based FS
DT	Decision Tree
ANN	Artificial Neural Network

PNN	Probabilistic Neural Network
NB	Naïve Bayes
RF	Random Forest
AB	AdaBoost
SVM	Support Vector Machine
LR	Logistic Regression
MLP	Multi-Layer Perceptron
SNP	Single nucleotide polymorphism
GA	Integer-coded genetic algorithm
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
RAE	Relative Absolute Error
RRSE	Root Relative Squared Error
CbH	Choice Based Hierarchies
ANFIS	Adaptive Neuro-Fuzzy Inference System
IFTS	Intent Fundamental Triage Scale
RF	Random Forest
ARFF	Associated File Format
UCI	UCI Machine Learning repository
CV	Cross Validation

ABSTRACT

Chronic Kidney Disease (CKD) is a chronic renal problem that affects the human kidney and makes it not to function properly or causes complete renal failure. It results in dialysis or causes other related diseases and reduces the quality of living. The symptoms of this disease cannot be identified in the preliminary stage. Only very lesser people are aware of this disease and can predict the symptoms at the earlier stage. However, it leads to prolonged disruption of kidney functional and finally causes it to failure and reduces the functionality completely. This can be occurred due to prolonged diabetes and also related with other diseases like Cardio-Vascular Disease (CVD). Due to inadequate prediction approaches lack of awareness in the preliminary stage, there is a delay in treating the patients' at the initial phase of disease. From the various literature studies, it is identified that CKD can be predicted and treated in the earlier stage using the soft-computational techniques. Earlier CKD predictor model needs to be improved with higher prediction accuracy and precision. Therefore, there is a need for a decision support system that assists the nephrologists during the time of emergency conditions. Therefore, in this research, an efficient Clinical Decision Support System (CDSS) is modeled based on patients data to identify the occurrence of CKD and Non-CKD with the expert decision support system. Here, Machine Learning algorithms are used for designing the CDSS to predict CKD in prior stage.

Initially, for analyzing the complications of this disease, data from UCI Machine Learning Repository for CKD is attained from online resources. Next, feature selection and classification is performed to analyze the functionality of the proposed model. Finally, performance metrics like accuracy, precision, sensitivity, specificity, Information Gain, Gini Index, and confusion matrix are analyzed for measuring the classes of the disease.

In the first phase of research, Naïve Bayes (NB) classifier is adopted for classification along with Choice based Hierarchy (NB-CbH). NB classifier works effectually with huge dataset and reduces the computational complexity. The prediction rate and the severity of the disease analysis with NB are extremely higher. With this prediction model, the dataset attributes are analyzed and the most dominant features are considered for further computation. The features like age, gender, eGFR, smoking, alcoholic conditions, BP and so on are considered as the dominant features. These

are given as input to NB-CbH model to measure the disease classes, i.e., (0-Non-CKD and 1-CKD). The prediction rate is measured with precision and classification error. The precision value based on Gain Ratio is 82.25%, with Information Gain is 91.5% and with Gini Index is 72.25% respectively. Similarly, the classification errors are 10.08, 9.09, and 22.35 respectively with the above-mentioned parameters. The overall precision is 88.7% and 90.2% for NB-CbH which is higher the other models.

In the second phase, feature selection is performed with Random Forest which can work effectually with the individual classifiers and reduces the error rate. Then, classification is performed with Adaptive Neuro-Fuzzy Inference System (ANFIS). This method is an intellectual classifier model to produce efficient prediction accuracy. Here, metrics like specificity, sensitivity, and precision are measured with 97%, 100% and 100% respectively.

The performance of the predictor model was analyzed with 10-fold cross validation method. The results obtained from ANFIS, and NB-CbH was interpreted with the performance measures such as precision, specificity, sensitivity, classification error respectively. The significant features identified by the predictor model are Age, gender, smoking, alcohol, and eGFR respectively. Therefore, the proposed classifier model gives higher rate of prediction accuracy for CKD in earlier stage. Simulation has been done with MATLAB environment where the proposed model shows better trade-off to other existing methods.

CHAPTER 1

INTRODUCTION

1.1. Prologue

Chronic kidney disease (CKD) is observed as one among a quickly growing non-growing diseases that induce the death rate significantly and sickness issues. In 2019, the statistical report states that 755 million people all over the world in which approximately 418 million were females and 337 million were males. It is extremely significant health care issues in India with 17.5% of world wide population. CKD is a condition in which kidney fails progressively to carry out functionality such as blood filtration (Sujatha et al., 2016). When this harm to the kidney occurs in a consistent manner over an excess time period, it is known as 'chronic' disease. The urinary organs are highly exaggerated or damaged. When this condition is worse, it leads to waste accretion in blood and affects the body which causes various health problems. Various symptoms are related to this chronic condition like hypertension, bone weakness, reduction in nerve damage, blood count level, and causes blood vessel and heart disease.

The prediction of CKD in prior stage and risk factor observation may helps in preventing further disease growth and limits the complication on individual's health condition (Yao et al., 2013). For measuring the severity, Glomerular filtration rate (GFR) is an out-standing kidney functionality measure. GFR determines how the kidney functions for the patient. The value decreases as the kidney condition gets worse. It is provided based on age, blood count, gender, race and added factors like sufferings of patients. It is classified into five levels or stages depending on GFR value. Various CKD stages with GFR level is provided in Table 1.1.

Table 1.1 CKD stages (clinical measure)

Stages	Parameters	GFR (mL/min)	Explanation
1	kidney function **(Regular)**	> = 90	No notable symptoms. Normal kidney

			function, however leads to other kidney diseases
2	Mild kidney damage	60–89	No notable symptoms. Kidney functionality is slightly reduced but leads to otherprobable kidney diseases.
3	Moderate kidney damage	30–59	Moderately reduces the kidney functionality. Patient experience high blood pressure, anemia and early bone diseases.
4	Severe kidney damage	15–29	Severe damage to kidney. Probability of kidney transplant is very high.
5	Established kidney failure	<= 15	Suffers from End Stage Renal Disease (ESRD). Kidney completely lost the ability to function and transplant/dialysis is required for survival

The measurement of CKD stages can be effectually analyzed with Machine Learning algorithms and makes the physician to take better decision with those attained values (R.S. Michalski et al., 2013). Similarly, Clinical Decision Support System (CDSS) helps physicians to take necessary actions during the time of emergency. The progression of CKD is shown in Fig 1.1.

Figure 1.1 CKD progression

The preliminary way to demonstrate kidney disease stages over the individuals are estimated using GFR (Davis et al., 2010). The following are the symptoms related to CKD:

- No proper sleep at night
- Loss of appetite
- Skin turns to be more dry and itchy
- Muscle cramps during night times
- Lack of fatigue and energy
- Loss of concentration on work
- Ankles and feet are swollen

1.2. Causes of CKD

CKD disease is due to the loss of kidney functionality and also due to the effect of various other primary conditions like hypertension, diabetes, and cardiovascular disease. In some urban countries, CKD are measured due to the outcome of hypertension. Some other factors like obesity, diabetes, old age, and cardiovascular disease (CVD) are accountable for the causes of CKD. High blood pressure and diabetes are some major causes of CKD.

1.3. CKD diagnosis

Renal function is considered for serum creatinine level measurement in the clinical medicine. GFR is evaluated with the creatinine measurement with serum creatinine levels based on urine samples in a timely manner. However, serum creatinine is generally utilized for evaluating the creatinine clearance (CC) with CKD, however serum creatinine is a poor GFR predictor and shows random production and influenced by various other parameters like body weight, sex, age, diet, drugs, and muscle mass. GFR is considered as a 'Gold Standard' for evaluating kidney function. The equation used for this measurement is "Cockroft-Gault (CG) equation" and MDRD equation. Some clinicians apply MDRD for its easier way of internet usage where values are added based on sex, race, weight, and age to identify GFR. This model shows huge errors and time-consuming factors. There are enormous equations that are formulated for eGFR measurement and the equations used for these measurements are given below. The techniques for evaluating GFR in ml/min/1.73 m are given below:

Cockcroft-Gault formula (CG)
Creatinine clearance (CC) = (140-age) (weight in kg) * (0.85) Serum creatinine (imol/l) * 0.81
MDRD
GFR = 186.3 * (Serum creatinine) – 1.154 * age – 0.203 * 0.742

1.4. Diagnosis Platforms

Chronic diseases (CD) are the foremost cause of loss of life and disability. The appropriate and suitable prediction of certain biomarkers of this disease reduces the infection and stops the death rate associated with this disease (McCormick et al., 2011). It is a normal clinical condition and considered as community health problem. The laboratory medicine functionality in CKD analysis shows huge significance.

There are diverse hospital-based and kit-based detection techniques available for the kidney disease prediction or renal disease prediction in patients. Some methods for predicting the CKD are:

- Albuminuria
- sediment abnormalities
- Electrolyte and tubular disorders
- Abnormalities detected
- Structural abnormalities
- kidney transplantation history

The above-mentioned predictions are hospital-based method which is time-consuming, not-specific, and costly for CKD. However, there are some kit based techniques over the market for predicting KIM-1, CysC for renal biomarkers for evaluating kidney diseases such as:

- ELISA for KIM-1
- Laminar-flow dipstick assay for KIM-1
- PENIA
- PETIA

Additionally, some gel-based detection kits are also accessible for CysC detection to evaluate kidney disease in earlier stage. The easier and earlier diagnosis with sensor-based platform for estimation and detection of kidney disease such as SPRI sensor that predicts cystatin in blood plasma 0.09 μg mL-1; however, the development platform is not so specific.

1.5.Predictive analysis with CDSS

The clinical decision support system (CDSS) helps physicians and other health care professionals to provide a proper decision-making process. Robert Hayward's states that, "CDSS combines health observation with health knowledge to improve healthcare choices"(Jin et al., 2018). CDSS makes use of Computational Intelligence (C.T.) techniques to make proper decisions. The decision support system considers various patients' data for the decision-making process. CDSS may significantly determine its functionality in the treatment phase (Weissmann et al., 2019). It is specifically used as an alarming or alerting system. It is utilized for providing

appropriate information to the individuals at the proper time with technology-based supporting automation. The ultimate role of CDSS is to provide alerting and sensing (Farran et al., 2020). Thus, CDSS is widely used in the diagnosis phase, as there are various advantages encountered in the machine CI field. The following are the applications related to CDSS:

1) Diseases diagnosis
2) Disease staging
3) Tracking of disease progression
4) Recommending effectual treatment process
5) Evaluating the results of the treatment process
6) Evaluating surgery results
7) Pharmacological decision support
8) Identifying the disease causality
9) Suggesting diet needs for patients
10) Prior warning systems

The efficiency of CDSS is determined by dealing with the challenges due to handling huge patients' data and data analytics. Machine learning algorithms along with CDSS aids to handle all the challenges, to resolve the issues and it can offer better accuracy in prediction. The general block diagram of CDSS is given in Fig 1.2

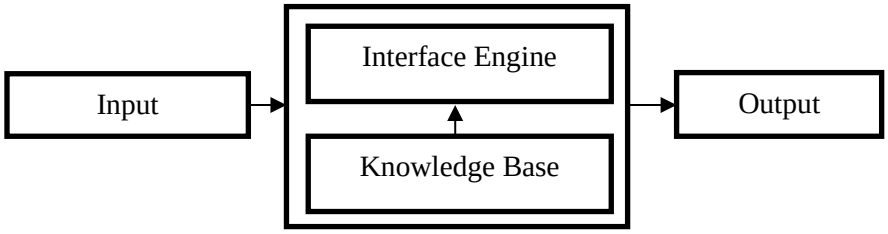

Figure 1.2 General Block Diagram of CDSS

The Clinicians interact with the CDSS to get suggestions for diagnosis while treating the patients. The Physicians accept the outcome of CDSS for treating RA in three diverse stages namely (i) Pre-diagnosis (ii) at a time of diagnosis (iii) Post-diagnosis as shown in Table 1.2.

Table 1.2 Stages of Clinical Decision Support System

Stage of CDSS	Advantage
Pre-diagnosis	Helps to prepare the diagnosis
During Diagnosis	Helps the Physician to filter preliminary diagnostic to improve further.
Post-diagnosis	Helps the Physician to extract the data and to derive relationship among patients and medical history and clinical data to predict events.

1.6. Benefits of using CDSS

The potential benefits of using CDSS are:

- Helps in improving test ordering and patient safety by reducing medication errors.
- Enhances quality of service by providing quality time for patient's care by facilitating up-to-date clinical evidence, guidelines and patient satisfaction.
- Reduces the cost of unnecessary laboratory investigations.

1.7. Machine Learning (ML)

It comes under Artificial Intelligence (AI). The ML functionality relies over the competency of the machine to address the problem devoid of certain computer program. The extensive ML applications over the medical field assist to endorse medical prediction, enhance the medical quality, and diminish the costs (Demsar et al., 2006). Moreover, various related investigations help to address the clinical issues with ML in nephrology departments which are still an essential need.

By having an extensive insight in the machine learning domain and thereby applying it in nephrology is the preliminary need to appropriately resolve and eliminate ML challenges. The computers have the competency to identify, learn, and judge human health conditions. It is essential based on the deployment and the development of diverse algorithms that applies statistical tools for describing the ML nature (Rish et al., 2001). This technology is extensively partitioned into unsupervised, supervised, reinforcement and semi-supervised learning model based on the modeling requirements. The division of ML is given in Fig 1.3 and ML techniques are given in Fig 1.4.

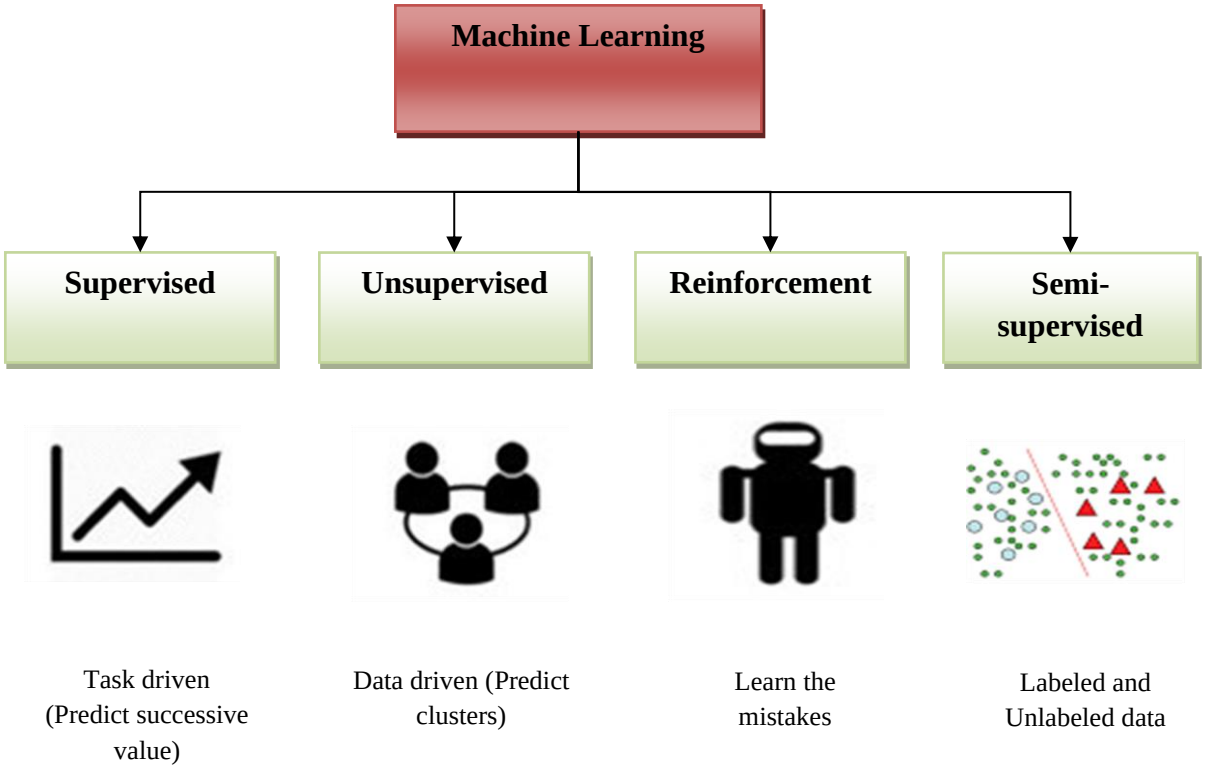

Fig 1.3 Types of Machine learning approaches

1.7.1. Supervised learning

The first learning method, i.e., supervised learning includes techniques like NB, RF, LR, and SVM is some of common ML form that are applied in medical research. Every instance of this learning model includes input object (generally vectors) and some appropriate output values (supervised signal). The applications of this learning process are extremely extensive (Forssen et

al.,2017. Moreover, there are certain constraints towards the applications as it shows complex optimal control parameters.

1.7.2. Unsupervised learning

When the learning instances are observed not to give essential information, then it is known as unsupervised learning. Alike of k-means clustering, this model partitions the samples optimally to various categories based in the training data characteristics without proper labels. Additionally, the unsupervised learning model attains intrinsic patterns over the histological data that plays crucial role in pathological prediction (Forssen et al., 2017). However, unsupervised learning pretends to build narrow gap among the AI and human intelligence.

1.7.3. Reinforcement learning (RL)

It addresses and describes agent problems by maximizing the returns and attain certain goal via learning process while interaction with corresponding environment. Here, Markov decision is considered as a general reinforcement learning model which captures the uncertainty related to the treatment process and the underlying random process (Aneja et al., 2014). This is specifically well-established for decision-making process sequentially where the issues like dosing for chronic diseases to predict dosage sequences.

1.7.4. Transfer learning

Transfer learning is also utilized to examine the smaller dataset. which has substantial effects on the performance of the model. The key advantage relies over the utilization of trained network model indeed of training or designing a newer network model. With this, knowledge and parameters are transferred. It consumes lesser amount of computational resources and execution time is needed to accomplish the newer tasks (Lundin et al., 1999). There is a need to complete the similar task, the appropriate model with pre-trained model is utilized to carry out transfer learning.

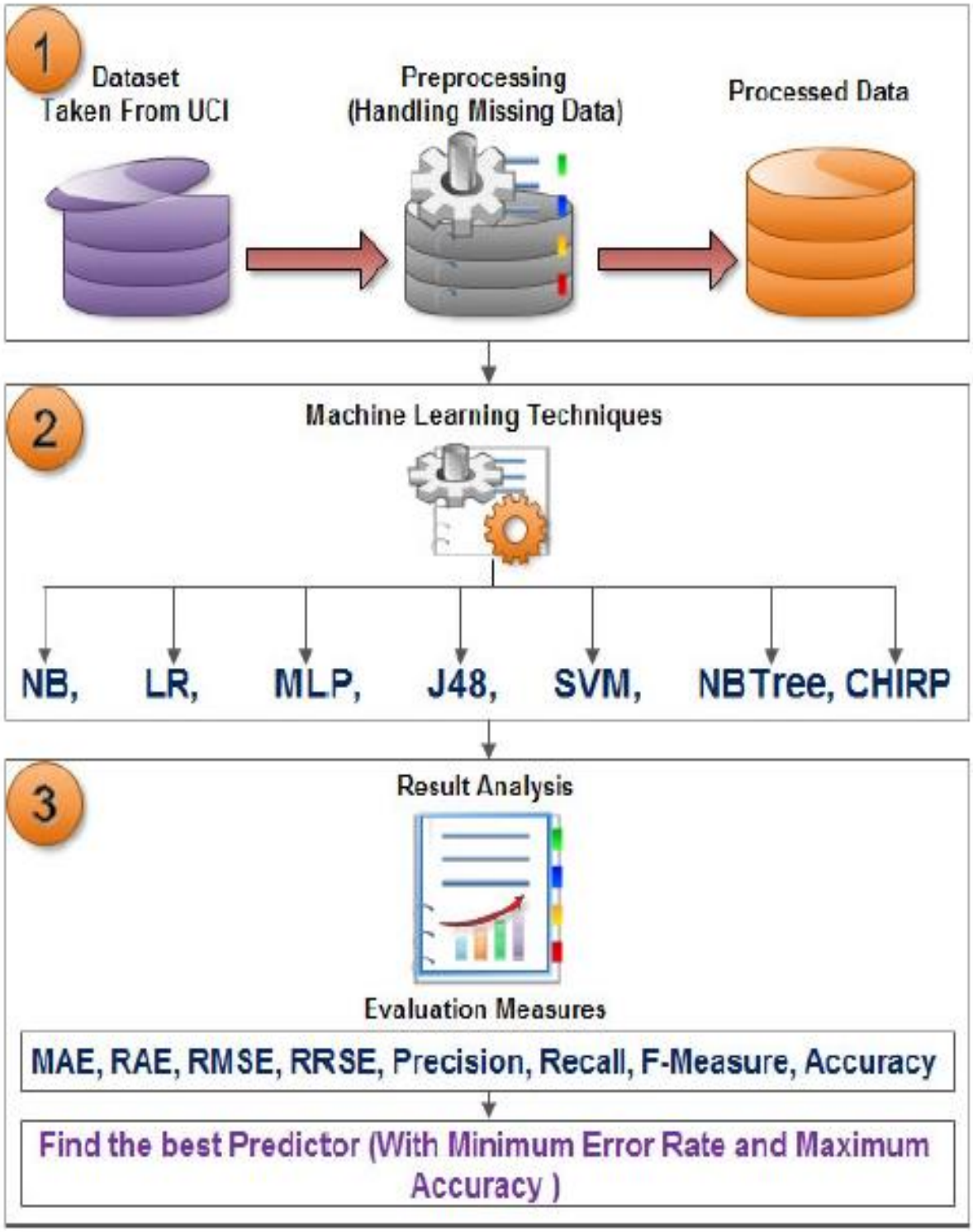

Fig 1.4 Machine learning techniques

1.8. ML for nephrology

Clinical data are considered as important resource. With constant enhancement of digital data over health care factors and AI technological improvements, ML is merged with the clinical data for prediction, prognosis and various related factors. In past few decades, various ML approaches are anticipated and used over the computational and medical biology (Khamparia et al., 2013). Additionally, the treatment and diagnosis process of the clinical disease by the electronic health records (EHR) which works effectually in the medical image analysis and genomics fields. Similarly, ML in cardiovascular disease field, the constrained evidence and limited research scope in predicting the kidney disease paved way that nephrology is not considerably beneficial with clinical application of ML.

Recently, the applications of precision medicine have made a huge progression towards the nephrology field. Nephrotic Syndrome Study Network (NEPTUNE) is executing precision medicine idea to design a disease definition with the multi-level and comprehensive disease progression analysis in cohort studies. The integration of big data and ML are considered as an essential factor in enhancing precision medicine. Even though, it is infancy, there are huge ideas and future for risk prediction in kidney disease.

1.8.1. Renal pathology

An ML application over image examination is extremely influential and showsfaster growth to validate the reliability of the method for examining the malignant tumors like cancer. In nephrology, biological image examination with ML is utilized for the renal pathology diagnosis with superior standard for renal disease prediction. The diagnostic process in turn influences the prognoses and treatment option series (Kunwar et al., 2016). Various prevailing techniques for glomerular evaluation is seem to be non-standardized, labor-intensive and manual. In recent times, to preserve the time and manpower, and to enhance the diagnosis accuracy devoid of any efforts and bias are seems to automatically quantify the glomerular injury.

1.8.2. Glomeruli and tubules segmentation

Segmentation is baseline concept of automated pathological prediction which appropriately specifies tubules and glomeruli structures from pathological images. In past studies, an unsupervised semi-automated concept is anticipated for segmentation and localization of these glomerular characteristics. Even though, localization accuracy has attained 87%, glomerular injury is not considered when the sizes are completely restricted to (148 glomeruli with 15 fields). The author applied supervised learning approaches from glomeruli prediction and establishes the model with 81% recall and 95% precision.

Moreover, it is validated that ML algorithm utilization not only merely segments glomeruli from kidney images;however differentiates renal tubule. SVM classification is applied for feature extraction over renal tubules in mice as depicted by Kunwar et al., (2016) where the True Positive Rate is 92%, False Positive Rate is 10%. With 200 cores of glomerular segmentation has ability to segment full-sized kidney of mouse in 40 minutes roughly. This is followed by the analysis of huge glomeruli than the manual execution. In addition, the SVM utilization predicts the variations in tubules and glomeruli in wild-type genotypic mice and knock-out where the pathological variation scores of mesangial matrix expansion and tubules vacuolation degree. The ML competency to predict and measure certain characteristics of renal tissue structures, more precisely, to perform renal tissue segmentation and appropriate scoring and assists in predicting the newer histopathological features is explained in detail.

1.8.3. Clinical factors

Some author used analytical approaches on pathological images. The patient specific trichromatic images are modeled and trained with input and clinical an indicator that includes the stages of CKD, nephrotic-range proteinuria, and serum creatinine during the biopsy period and 1-, 3-, and 5- years of survival which is an outcome to this investigation. The diverse CNN were trained and performance is compared with prevailing models by experienced to identify CKD stages with 0.519 and 0.051 kappa-values correspondingly. AUC is 0.912 for CNN and 0.840 for

fibrosis score over creatinine measure and 0.867 for CNN and 0.702 for PEFS measures. CNN based AUC value of 1-, 3-, and 5- years survival is measured as 0.878, 0.876, and 0.905 where AUC values for these models are 0.812, 0.801, and 0.787. These outcome demonstrates that the clinical and effectiveness feasibility of applying this model in renal pathology. This identification validate the additional values for biopsy outcomes with other clinical evaluation, and offers more appropriate follow-up strategies and care management for those biopsy patients. Moreover, no beneficial validation is performed till now.

1.9. Kidney diseases

With initiation of ECHIT act, EHRs are drastically increasing. EHR data holds essential information regarding the disease evolution. With the processing of ML, it is extensively applied in disease prediction. This provides potential benefits to predict and recognize the renal disease progression and renal functionality damage. Based on the EHR data, it is probable to acquire more effectual understanding towards the patient's health status and appropriate prediction of risk factors to acquire certain diseases (Pujari et al., 2014). Recently, the big data era and the technological advancements with diverse ML algorithm examine the standardized health information for routine collection and to perform large-scale observation research which is extremely popular and crucial. The model prediction is provided with HER data is considered to provide individualized prediction of kidney diseases and to enhance medical treatment quality. The risk factors associated with CKD is shown in Fig 1.5.

1.10. Prognosis and diagnosis of kidney diseases

It is found that ML is extremely likely to identify eGFR more precisely. Various studies determine the significance of the AI and extremely likely to recognize the progression of diverse CKD more accurately. Tangri et al., () uses laboratory data like albuminuria and eGFR from the routing EHR information and establishes the extremely high proportional hazard model for CKD with renal failure and executed certain validations, i.e., C-statistics 0.922.

As well, it is validated that temporal information can enhance the prediction capability with renal worsening. The consideration of temporal information with medical data can recognizes the renal functionality loss and predicts higher risk with short-term renal failure.

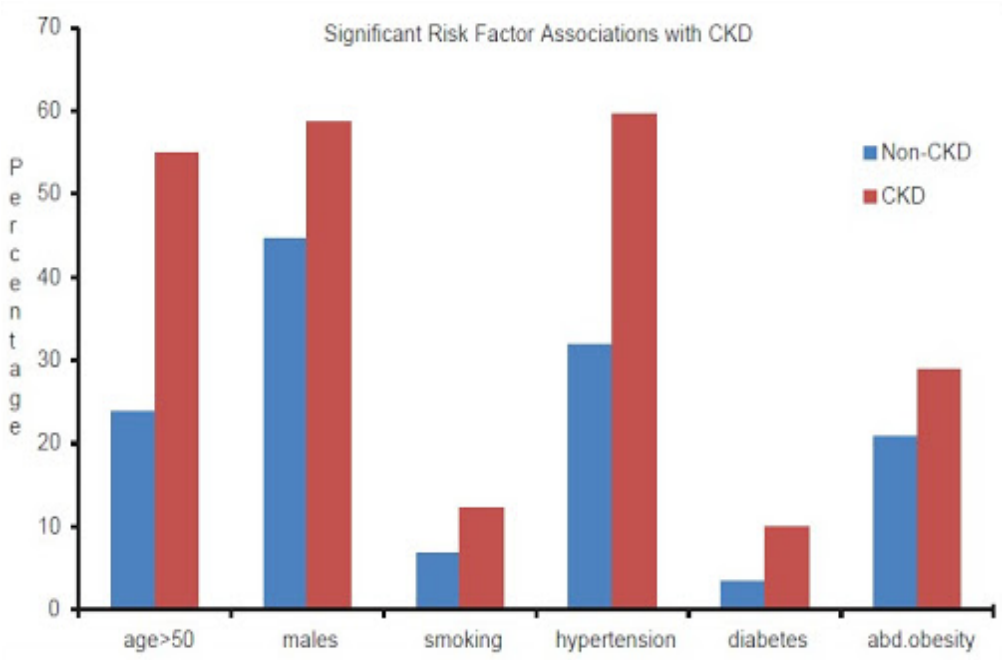

Fig 1.5 Risk factors associated with CKD

With constant enhancements in AI technology, huge advanced ML algorithms and more complex algorithms are applied and explored in modeling. The author considers ten years clinical information for predicting CKD and use of adaptive neuro-fuzzy interference system to analyze timeframe of renal failure with CKD. This model evaluates the eGFR variations more accurately for all the consecutive periods (normalized MAE < 5%). As well, with the application of diverse supervised learning algorithm, The author applied unsuperived learning model known as latent Dirichlet allocation to execute clinical data and determines prediction model for CKD from stage III to stage IV.

At present, there are diverse limitations that prevail and not yet resolved. Even though, highly performing risk prediction models are established increasingly, appropriate model effects still has to be explored more. After successive modeling, it is essential to finely validate and calibrate the outcomes externally and to verify consequences of assessment results before merging

it with the guidance principle (Wibawa et al., 2017). The research advancements of two diverse kinds of CKD, immunoglobulin A nephropathy (IgAN), diabetic kidney diseases (DKDs) are discussed below.

1.11. Acute kidney injury

In various countries, the mortality rate towards the acute kidney disease is 10% - 12%. There are huge efforts are given to the clinical results of AKI which have been concentrated on earlier prediction and personalized treatment. The earlier evaluation of AKI reduces the mortality rate and the renal prognosis can be improved considerably (Adam et al., 2012). Various models are provided to realize prior prediction of disease with monitoring based on real-data objectives; however, it saves the energy and time of nephrologists. With AKI determination with clinical practice guidelines and essential EHR application growth in big data field, an enormous amount of EHR data initiated to play crucial role in AKI-based clinical researches. Now, it is extremely essential tool for diagnosing and predicting AKI. The evolution of CDSS which relies on self-learning predictive model is applied over the AKI hospital environment for monitoring in the future clinical practices. The Generic view of affected and normal kidney is depicted in Fig 1.6.

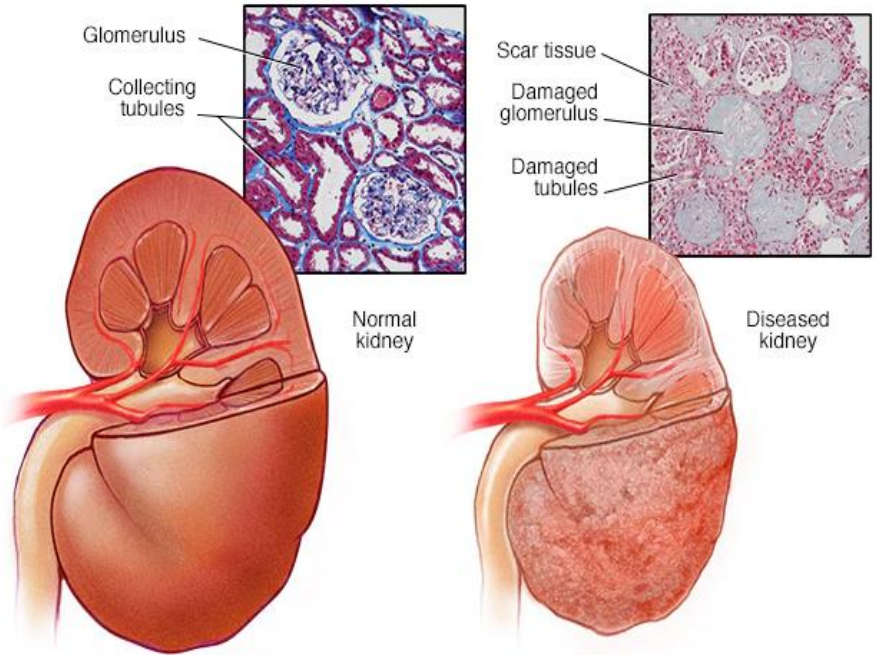

Fig 1.6 Generic view of Normal and diseased kidney

1.12. AKI earlier assessment

The prior or earlier AKI assessment decreases the mortality rate and enhances renal prognosis. GBM is considered as a AKI prediction model after cardiac surgery and liver transplantation using various ML algorithms. GBM provides superior performance, and AUC over liver transplantation is associated with 90%, while other gives 78%. In addition, Huang et al., (2009) provides prediction approach for AKI after PCI that relies on GBM. The prior studies include huge amount of data from enormous patients undergoes PCI surgery to determine the baseline model (Gunarathe et al., 2017). As well, temporal validation is performed with the data attained from patients. AUC is measured to 78.5% which is superior to the LR baseline model (AUC 75.3%). Henceforth, the enhancements of algorithms and enormous data show the efficiency to offer appropriate risk estimation.

1.12.1. AKI prediction towards death risk

The death risk determination in AKI patients was considered and various approaches are modeled. With the utilization of medical information mart database, Lin et al., (2010) modeled mortality prediction approach with an accuracy of 72% and AUC of 0.866 for AKI patients using random forest. Moreover, this model overestimates patients' mortality slightly with lesser death risk and underrated mortality (Dua et al., 2017). With widen insights in recent research, various ML methods assist in monitoring the essential and useful outcomes of AKI and diminishes the mortality and morbidity. These inspiring advancements have to be validated in further research.

1.12.2. Dialytic treatments

ML algorithms are extensively utilized in dialysis monitoring and prescription, death prediction, complication management and so on. There are also huge potential for child dialysis application. The ML applications in dialysis have provided enormous positive research outcomes with dialytic treatments. The investigators have arranged the first scientific conference of the future AI development and the application status during AI dialysis. In this scientific conference,

the AL experiences during dialysis are completely analyzed and the challenges, obstacles and future field applications are also analyzed. The outcomes specify an extensive application with the superior prospect with AI during renal dialysis. In future, AI shows progressive variations in clinical practices based on hemodialysis. The members associated with Nephrology conference in global platform 2020 is shown in Fig 1.7.

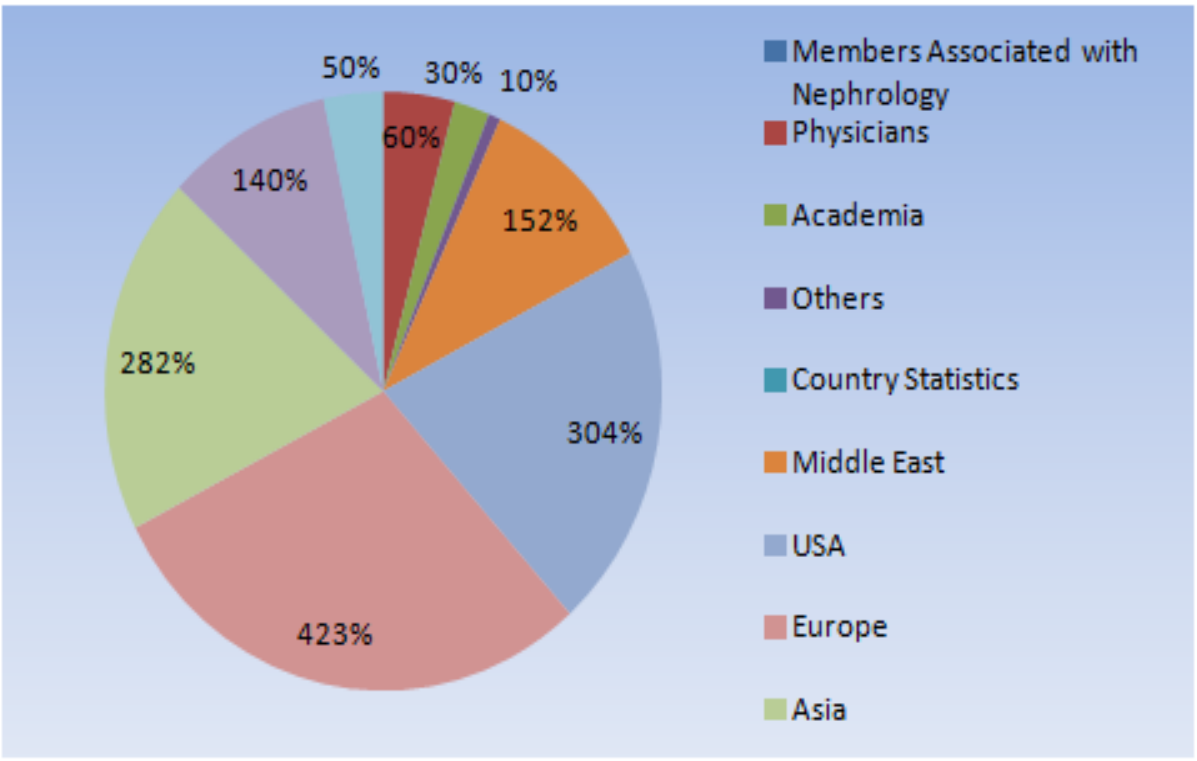

Fig 1.7 World Nephrology 2020 is a global platform for the Nephrologists

1.12.3. Death prediction

Some hemo-dialysis patients face death as an outcome of diverse complication after or during hemo-dialysis. It shows higher mortality rate during the first year of dialysis. The accurate computation with post-dialysis mortality assists clinicians and patients to make better decision possibly. Dua et al., (2017) applies RF algorithm for construction of diverse mortality prediction approach with huge patients after transition dialysis. C-statistics is 0.7186, 0.7444, 0.7505, and 0.7480 which offers finest internal replication and effectiveness. It accurately analyze post dialysis patients mortality, however there are some drawbacks with the external validation.

Some deaths are sudden due to severe cardiovascular complication during dialysis. ML approaches are provided to help the estimation with sudden cardiac death prediction. With the preliminary patients' information after and before dialysis, RF prediction is developed by Dua et al., (2017). C-statistic is considered to be 0.799. As competency to identify death to decreased prediction time where data gathered are extremely conductive to evaluate short-term risks than the long-term.

1.13. Challenges

Even with ML development, there are still some necessary factors that have to be improved constantly to fulfill the challenges in the traditional clinical practice. Based on these limitations with ML, the intrinsic logic behind ML models are alike of black box which is extremely complex to be examined by the Doctors.

In addition, the ML ethics of ML needs to be considered. Even though, there is certain guideline that needs to be emerged, Artificial Intelligence Governance provides guideline to various private sectors over the use of AI algorithms; the improvements in obvious clinical guidelines are lagging behind AI technological progression (Dua et al., 2017). It is essential to model a guideline as earlier as possible to establish the standardization of clinical application. Even though ML model performs well during training process, stability is needed to determine the regulatory and privacy requirements with application of diverse and huge datasets for enhancing accuracy of ML models. There are diverse challenging factors that are related to the CK and renal disease/failure which is discussed below.

1.13.1. Challenges for nephrology

Along with the constraints in the advancements of ML algorithms, there are diverse challenging applications over ML in nephrology.

1.13.2. Challenges in clinical data processing

During the data collection process, it is extremely complex to gather the appropriate EMR and various medical institutions lags in standardization and uneven, missing or incomplete data. This outcome in lower data quality and complexity is hauling out the essential information. It is essential to fulfill authenticity and integrity of data collection. The most general way in expressing the limitations in diverse studies are the lacking in external verification. Some of the investigations are performed by considering the outdoor model verification to authenticate the viability and precision of the given model. Also, it is completely related to the complexity in data acquisition. Recently, data over every unit are non-shared and scattered basically (Adem et al., 2019).Some data are confusing, larger and complex. It has to be appropriately mapped before it is applied for designing. This is a basic step for model as accuracy is completely based on data reliability in clinical reflection. Indeed of projecting how much effort given to the enhancement of ML algorithms, any sort of inaccuracy in label is seriously constrained the exploitation of ML algorithmic accuracy. Thus, it is extremely challenging to consider and to acquire the high-quality data and makes it more accessible (Ahmed et al., 2014). In addition, the cost of modeling process is also to be considered. It takes enormous amount of material resources and manpower to gather data and to store it over the computer to process using large-scale computing power. Before model initiation, investments of funds are also challenging.

1.13.3. Challenge towards pathological diagnosis

Various researches concentrate on stage of appropriate prediction of glomeruli. Sample data size is considerably smaller, and images utilized for designing are generally from pathological kidney section. Even though, AI growth is tremendous, owing to its certain complexity in pathological demonstration of diverse renal disease and nearer association shows clinical indicators, automatic pathological prediction of certain renal diseases has not yet analyzed (Ahmed et al., 2014). It is extremely probable to substitute pathologists completely for pathological prediction of types of kidney diseases via comprehensive patient data. It is assisted with huge amount of data and validates the extensive studies. The modeled framework functionalities over

diverse clinical practices and datasets are validated by these approaches in distribution and lesions spectrum identified in typical renal pathological services.

1.14. Problem statement

There are major problems that are associated with Chronic Renal disease, despite of Acute Renal failure which is developed over various years and months. Similarly, the medical institutions do not hold any certain scale for measuring the probability of this chronic renal disease in the primary or in the intermediate stage. Various hospital information systems are modeled to assist inventory management, patient billing, and production of simple statistics. Some hospitals apply Clinical Decision Support System (CDSS); however they are highly constrained. Some simple queries such as "What is the average patients' age who is identified with CKD"?. "How many surgeries or transplantation shows positive results in hospital regarding the CKD treatment process?". "Predicting the age of patients who have the possibility and high risk of CKD?" Similarly, some complex queries are not answered such as "Recognizing the significance of pre-operative predictors that increases the duration of staying in hospital?". How many patients are treated with dialysis and their pre-operative and post-operative symptoms? With the available patients' record, prediction of patients with the probability of higher rate to cardiovascular disease.

Conventionally, Doctors does not possess a clear way for predicting with the fact that their patients are headed to Chronic Kidney Disease (CKD). Most patients' does not have any clue to predict the certain causes of renal abnormalities. Generally, physician's needs to take appropriate decision based on the expertise knowledge, indeed on the huge knowledge with rich data with missing data rate over the dataset. This leads to unnecessary errors, biases, and excess medical costs that influence the quality of service given to the patients. Therefore, there is necessity to integrate clinical decision support with patients' database for intellectual decision making. With this context, Machine learning is a superior approach to model a knowledge rich environment to attain superior clinical decisions.

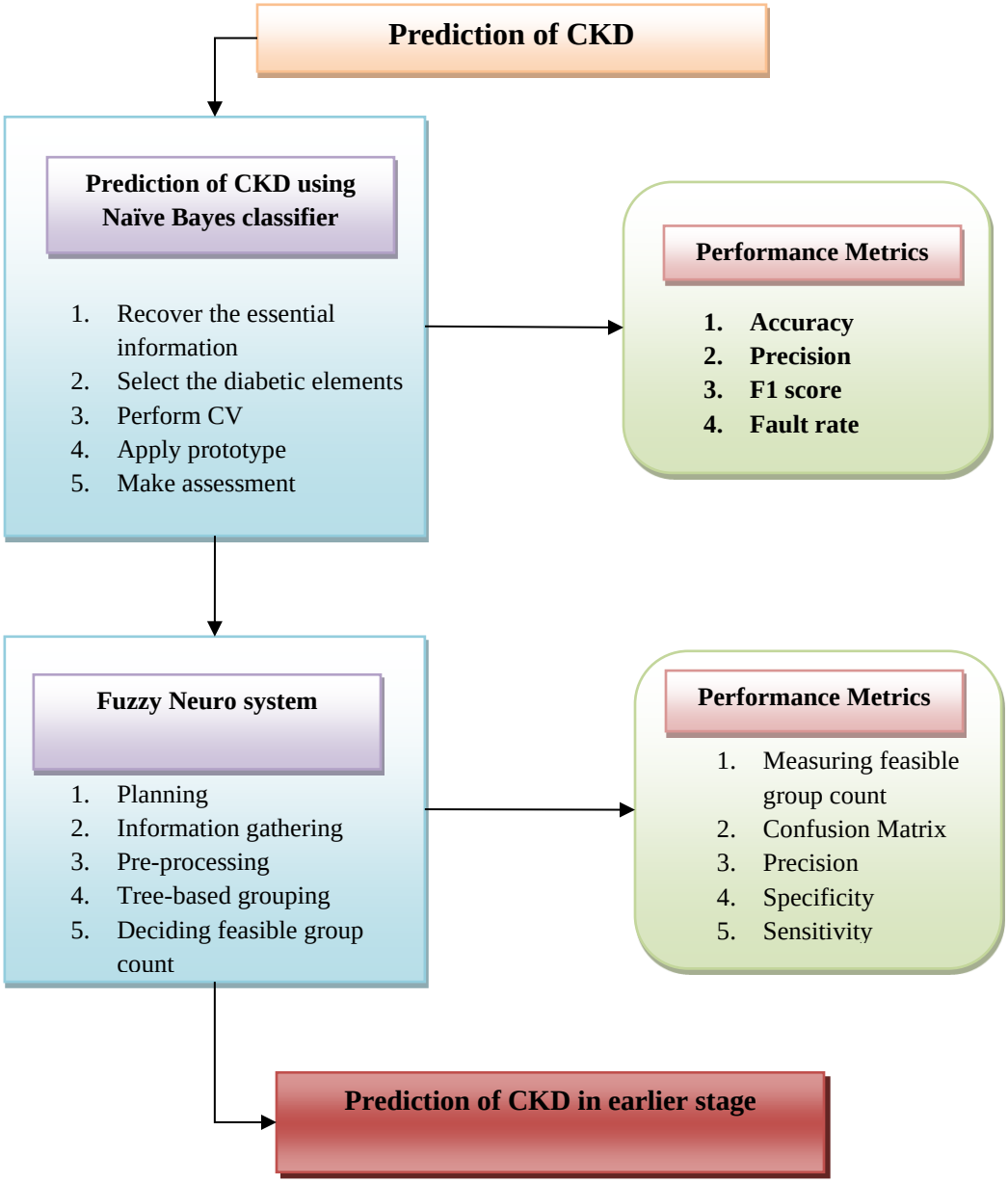

Fig 1.8 Research methodology

1.15. Motivation

Machine Learning plays an indispensable and snowballing role in medical field applications by assisting physicians to make prior prediction process and to reduce the death rate. Similarly in nephrology, ML algorithms are applied to identify the progression and the disease risk along with the hemodialysis analysis and follow-up. Even though, it is considered to be a scratching surface, the technological progression, data accumulation and huge investments, this ML technology makes

a major breakthrough in the nephrology field. Thus, to providing a solution by identify the disease severity in the earlier stage and to treat the patients for reducing the mortality rate. The reducing in mortality rate and with the enhancement in treatment process motivates the application of Machine Learning algorithm in predicting CKD. Fig 1.8 depicts the research flow of this investigation.

1.16. Research objectives

The following are the essential research objectives concerned with the prediction of CKD:

- To make an initial prediction strategy towards CKD for diabetic patients using Naïve Bayes approaches.
- To generate the prediction metric like accuracy, sensitivity, and precision based on the model requirements.
- To evaluate the functionality of proposed model with other prevailing methods.
- To apply neural fuzzy approach for predicting the severity of CKD with other consequences.
- To measure the data that influences the chance of CKD in earlier stage and to take preliminary actions to avoid the kidney damage.

1.17. Research Scope

The scope of the research is to extend the evidence for CKD assessment with predictor model. This model provides clinical experience to Physicians for evaluating prognosis in Patients. Further, it also maintains feasibility for assessment with the available resources in proper time scales. The validation or impact of eligible outcome helps to predict treatment response in CKD patients. The final scope of this study is to bring out the impressions of the patients who are diagnosed with CKD according to established criteria. This model works well even in case of least available resources.

1.18.Thesis organization

The thesis is structured as:

Chapter 1 elaborates CKD disease, prediction analysis with CDSS, Machine learning approaches, ML in nephrology, kidney diseases, prognosis of kidney disease, acute kidney injury, dialytic treatments, challenges, problem statement, motivation, research objectives, and thesis organization.

Chapter 2 depicts the reviews on CKD pre-processing, data processing, predictor model, classification and prediction techniques in bio-medical field, assessment criteria based on CKD disease prediction

Chapter 3 explains about the novel Naïve Bayes classifier used for chronic kidney risk prediction using Machine Learning schemes. The precision measures are considered as the evaluation metrics to validate the functionality of the algorithm over earlier prediction of CKD.

Chapter 4 describes the threat level prediction for CKD using the Neuro-Fuzzy schemes. Here, metrics like sensitivity, precision, specificity are considered along with confusion matrix. This model is applied for predicting the severe kidney ailments.

Chapter 5 explains the numerical results attained by evaluating Naïve Bayes classifier, and Neuro-Fuzzy system for the prediction of CKD.

Chapter 6 is the conclusion/summary of this research with the future research directions for the young and upcoming investigators.

CHAPTER 2

LITERATURE REVIEW

2.1. Prologue

This section discusses in detail about the various reviews and analysis performed by the investigators to predict CKD in earlier stage and the drawbacks related to it. CKD is a crucial medial disease that influences 10-15% of the population of the world and the occurrence is constantly increasing. CKD is not predicting in its earlier stage. An individual with CKD shows higher probability of heart disease development. The preliminary CKD stage does not any huge symptoms and it is extremely complex to predict it devoid of any tests such as blood test or urine test. When it is predicted in the earlier stage, better treatments and preventive actions are provided to have a control over the disease and to avoid transplantation and dialysis. Various studies by the investigators reported the prior detection of CKD which can diminish the disease growth by the Doctors and the nurses who are specialized in nephrology. In general, imaging techniques are used for predicting CKD. However, due to the huge amount of patients, it is completely unfeasible for testing person to person with better probability in CKD is suggested to experience testing. Presently, clinical database preservation turns to be complex process in healthcare industries and time consuming process.

Here, extensive reviews are done with the eGFR range, CKD in India, CDSS for CKD prediction, dataset analysis and pre-processing, predictor model, classification and detection techniques, research gaps which is discussed in the section given below.

2.2. eGFR reference range

Investigators like Soares et al., (2013) performed certain cross-sectional survey with set of people to evaluate the GFR reference range. The GFR measurements were performed using 51Cr-EDTA single-injection approach. The outcomes revealed that the mean of GFR computation is 106 ± 18 ml/min/1.73 m2. There is no appropriate variance that are encountered among female and male GFRs (109 ± 19 vs. 105 ± 19 ml/min/1.74 m2, P = 0.135 respectively). The age of the individuals should be ≥45 where the GFR values are lesser while compared with the individual

with an age less than 45 years (99 ± 16 vs.113 ± 19 ml/min/1.74 m2, P < 0.001). With this mean value ± 2 SD, GFR reference values is founded as 76 to 148 ml/min/1.73 m2 for the subjects < 45 years and 68-128 ml/min/1.74 m^2 for subjects higher than 46 years, irrespective of gender.

A cross-sectional analysis were done in India to evaluate the renal functionality using Diet modification in Renal Disease and Cockcroft – Gault equation for roughly of 490 healthy south Indian males with a target to evaluate the GFR measurement using MDRD and CG equation and to measure the relationship among the eGFR and body mass index (BMI). The outcomes demonstrates that the analysis give mean value of 91.05 (± 15.04) eGFR value and and 86.43 (± 13.61) value with CG and MDRD respectively.

Similarly, an extensive analysis is done in India with a target to demonstrate the GFR reference range with the healthy Indian adult kidney donors. GFR measurement for 610 subjects (360 females, 250 male, average age 35 years) healthy kidney donors with 99mTc-DTPA (diethylene triamine pentaacetic acid) are analyzed with two-plasma sample techniques. The outcomes demonstrate that mean GFR value of adult kidney donor was 4 ± 19.5 ml/min/1.74 m^2 BSA, i.e. diverse from normal value of 110 -126 ml/min derived significantly.

2.3. Chronic Kidney Disease in India

CKD is considered to be one of the decisive health problem and it raises indisposition, mortality, and health care expenditure. Diabetes Mellitus and Hypertension are some common issues related to CKD. eGFR is a measure of kidney function. There are five different stages in CKD with eGFR. The final CKD stage is termed as ESRD and with this stage patient requires kidney transplantation or dialysis. CKD prediction in prior stage assist in delaying disease progression which reduces economic burden on patients', community, and family. CKD prediction is earlier stage is also done with CKD- EPI (Chronic Kidney Disease - Epidemiology) formula. Prior prediction also assists in preventing the complications by referring patients to the nephrologists.

a) CKD prevalence

The number of individuals with chronic kidney disease treated by transplantation and dialysis has augmented. Every year 1, 00,000 individuals are diagnosed patients of ESRD to initiate dialysis in India. From India, constraint data are accessible for CKD prevalence. In India, hypertension and diabetes for 40-60% CKD cases are identified. Diabetes prevalence in adult population (Indian) was 13.6% in Chandigarh and 5.3% in Jharkhand. The occurrence of diabetes mellitus in Thiruvallur and Kancipuram district, Tamil Nadu was 10.4%. With rising prevalence of diseases in India, CKD prevalence is considered to increase, and clearly this is the primary target population to resolve this issue. CKD screening investigation has to be performed in India to sensitive people regarding CKD stages and educating population regarding hypertension and diabetes management to eliminate complications.

2.4. Computer Assisted Decision Making System

CDSS is more practical with time-sharing operating systems, computers, and distributing computing generation. The system implementation begins in mid-1960s. In technological field, chronicling history is neither neat nor linear. A diverse person recognizes DSS field from diverse vantage points and report various accounts. Technological evolved newer computerized DSS applications were studied and developed. Fig 2.1 depicts the pictorial representation of CDSS.

Investigations utilized diverse frameworks to construct and understand systems. The number of diverse approaches are organized to automated decision-making systems are explained. The issues of offering computational support for medical diagnosis are approached from various directions such as fuzzy logic, logical reasoning, set theory, Bayesian networks, if-then rules, traditional parametric and non-parametric statistics, case-based reasoning, case-based reasoning, possibility theory, support vector machines, and aggregations or technique combination. An extensive philosophies and systems are applied to medical diagnosis.

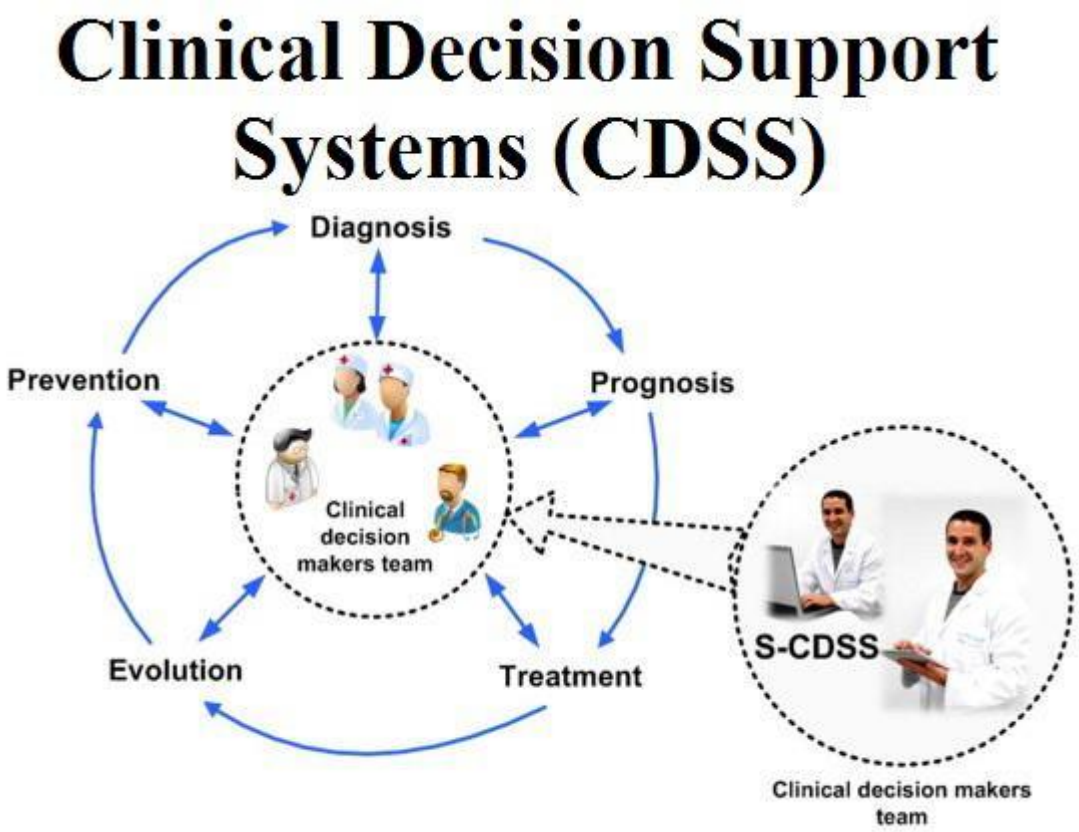

Fig 2.1 Clinical Decision Support System

2.5. Reviews on Decision Making Systems

The target of using computer based medical decision support system is to facilitate medical professional to utilize computer as a tool during decision making procedure. However, every physician are challenging during formation by the learning process for diagnosing. Here, the investigators need to resolve the deducing problem with some diseases or treatment formulation based on more specified observation and knowledge like heart disease or Hepatitis-B. Indeed, there is some certain knowledge of courses, seminars, and books; however, hand medical knowledge is out dated quickly and it does not substitute its own experience. For this functionality, some essential complexities are considered.

1) The basis for appropriate diagnosis, sufficient number of cases are reached at physicians' career and it is not present during the academic formation.

2) Especially, it is true for newer or rare diseases are in similar situations when entered as new comers.

3) In general, humans do not resemble static evaluations, however pattern recognition systems. A human are considered as objects or patterns which are extremely easier but fails during the probability of allocating the observations.

These principles are extremely complex and are not extensively known by the physicians. As well, the study does not provided that 50% of diagnosis is made wrong and do not obstruct self-consciences of certain physicians. A significant solution for these problems relies over the systematic applications of statistical instruments. The finest availability over computer ameliorates the probability of inexperienced physicians statistically to use the advantages of this kind of diagnosis:

1) During learning phase, physicians with lesser experience are attained with appropriate diagnosis using gathered data of experienced colleagues.

2) The disease condition is extremely complex to predict; for instance, Heart Disease, and Hepatitis-B, it is probable to attain superior prediction; if they utilize worldwide experience of networked colleagues.

3) Unknown diseases are documented systematically; even if this computation is complex which is not known to treat physicians.

4) As well, during the treatment of standardized diseases a crucial statistical evaluation for the application of operation techniques or therapies initiate’s doubts in physicians own, as it is generated by medicine proof.

2.6. Reviews on CKD pre-processing

The data is altered and cleaned before performing certain process by classification. Pre-processing stage comprises of missing value estimation, noisy data exclusion like normalization, outliers, and unbalanced data examination. Usually, real-world data is composed of various missing or inappropriate values. The easier way is to eliminate the complete records with missing value; however, it is not so appropriate for least datasets (Kunwar et al., 2016). The easier way to remove the records with missing values is data imputation process. Missing attributes are median imputed where it is replaced by missing values of mid-range attribute values. For supposed attributes, mode imputation is carried out while replacing the mislaid values with the most appropriate attribute values. The attributes in CKD dataset are showcased in Table 2.1.

Table 2.1 Attributes of CKD Dataset

S.No	Attribute	Class	Explanation	Ranges	Missing rate
1	Age	Numerical	Age	2…. 90	2.25%
2	BP	Numerical	BP	50… 180	3%
3	Sg	Nominal	Gravity	1005…. 1025	11.75%
4	Al	Nominal	Albumin	0… 5	11.5%
5	Su	Nominal	Sugar	0… 5	12.25%
6	Rbu	Nominal	RDC	(normal, abnormal)	38%
7	Pc	Nominal	Pus cells	(normal, abnormal)	16.25%
8	Pcc	Nominal	Pus cell clumps	(present, not present)	1%
9	Ba	Numerical	Bacteria	(present, not present)	1%
10	Bgr	Numerical	Blood glucose random	22…. 490	11%
11	Bu	Numerical	Blood area	1.5…. 322	4.75%

12	Su	Numerical	Serum creatinine	0.3…. 76	4.25%
13	Sod	Numerical	Sodium	4.5…. 150	21.75%
14	Pot	Numerical	Potassium	2.4…. 47	22%
15	Hemo	Numerical	Haemoglobin	3.1…. 17.8	13%
16	Pcv	Numerical	Packed cell volume	9…. 53	17.75%
17	Wc	Numerical	WBC count	65…. 26400	26.5%
18	Rc	Nominal	RBC count	2.1…. 8.5	32.75%
19	Htn	Nominal	Hypertension	(yes, no)	0.5%
20	Dm	Nominal	DM	(yes, no)	0.5%
21	Cad	Nominal	Coronary artery disease	(yes, no)	0.5%
22	Appet	Nominal	Appetite	(good, poor)	0.25%
23	Pe	Nominal	Pedal edema	(yes, no)	0.25%
24	Ane	Nominal	Anaemia	(yes, no)	0.25%
25	Class	Nominal	Class	(ckd, not ckd)	0%

2.7. Reviews on data processing

Every nominal (categorical) variable are coded to facilitate computer processing. For the pc and rbc values, abnormal and normal are coded as 0 and 1. For ba and pcc values, not present and present are coded as 0 and 1. For htn, cad, dm, ane, and pe are not coded as 0 and 1. For appet values, poor and good are coded as 0 and 1. Even though with original data descriptions determine three variables su, al, and sg as categorical types, the variable values are numerical. Therefore, these variables are considered as numerical variables (Khamparia et al., 2013). All these categorical variables are changed to these factors. Every sample is provided as an independent number ranges from 1 to 400. There are huge amount of dataset missing values, and number of complete samples are 160. Generally, patients' may miss certain measurements for diverse causes before diagnosis. Therefore, missing values are seems in the data; when diagnostic sample categories are unknown ad the imputation techniques are required.

After categorical variable encoding, missing values with original CKD dataset were filled and processed. K-NN imputations are used and it chooses ‘K’ complete instances with shorted ED for all samples using the missing values. Using these numerical values, missing values are composed with K-complete sample variables (median), and for category variables, missing values are completed with category with highest frequency with K complete samples (appropriate variables) as stated by Jin et al., 2014. For physiological measurements, individuals with same corporeal conditions possess physiological measurements using techniques like k-NN to fill mislaid values. For instance, Measurements which chiefly focuses on functionality of key organs of the human body are more likely to be unchanging with appropriate amount of hale and hearty individuals.

For every diseased individual, physiological measurements of the individuals with same degree of alike disease are same. Specifically, differences among the physiological data measurements are not large for people with same situations (Jin et al., 2014). This approach is adopted for diagnosing data and it is used in hyperuricemis area. When median variables with K samples are chosen, ‘K’ is rather considered as odd number; as the middle number is generally median; when numerical variables in ‘K’ entire samples are sorted with numerical values.

‘K’ selection is neither too larger nor too smaller. Excess ‘K’ larger values are ignored during non-noticeable mode which is more significant. On the other hand, an excess ‘K’ values leads to abnormal and noise data that influences the filling of missing values remarkably. Henceforth, ‘K’ values are selected as 3, 5, 7, 9, and 11 respectively. As an outcome, five entire CKD datasets are constructed. Additionally, this is proven with the effective k-NN imputation by evaluating the two other techniques. The first is to utilize random values are provided to fill the missing values; next is to utilize mode and mean of variables to fill missing values with categorical and continuous variables correspondingly.

2.8. Reviews on predictor model

Various approaches are anticipated for effectual CKD prediction by the utilization of patients’ medical values. Here, Cuckoo Search with neural network (CS-NN) techniques is

explained for predicting CKD. Firstly, the model is constructed to handle these crises that prevail over local search learning techniques. CS process assists to choose input weight vector optimally using NN for training data appropriately. Classifier outcomes of the model are demonstrated with superior performance (Jin et al., 2014). An improved version is modeled to get rid of local optima problem. When neurons' initial weights can manage NN performance, the anticipated model utilizes MCS algorithm to reduce RMSE values assigned during NN training process. The simulation outcomes are reported; when NN-MCS algorithm acquires superior functionality than NN-CS approaches.

The author discusses two fuzzy classifiers are termed as FOAM and FuRES are explained for CKD identification. FuRES produces classification tree neural network which is minimal. It produces classification rules to describe weight vector with lesser fuzzy entropy. 386 CKD patients' identification is done using two classifiers of fuzzy type. As well, FuRES is superior in contrary to FOAM where training and prediction process comprises of noise intensity. FOAM and FuRES acquired superior performance in CKD identification; similarly, FuRES is capable than FOAM. Subsequently, fuzzy-based techniques are accessible to predict CKD.

The author modeled an IHFCM, an enhanced version of FCM using ED for CKD detection. This work exploits probability based techniques are not appropriate for CKD prediction due to inevitability of appropriate output (Jan et al., 2014). Statistical techniques, association rule, Bayesian classification are infeasible to utilize inappropriate techniques. Therefore, IHFCM is generated for CKD identification. At preliminary stage, IHFCM eliminates frequent rows in the pre-processing step. It evaluates diffuse score individually in specific table of query contents. The fuzzy score specifies higher risk clusters and lesser fuzzy score specifies no risk or lower.

Giannouli et al., (2019) modeled a newer approach termed as WAELI. Missing value diminishes CKD precision level. As prevailing techniques, the adoption of data pre-processing approaches, data cleaning is essential to deal with missing values and eliminate inappropriate values. A re-calculation process is provided in diverse CKD stages in which the values are evaluated. Even though, the prevailing techniques are more effectual; it requires an expert to fulfill the CKD values.

FS procedure works as an effectual process during data classification utilized to determine rule-set which are small from the vast training dataset with appropriate outcomes. Various techniques like bio-inspired algorithms, AL techniques are utilized in FS. A technique with wrapper is offered by GA hybridization with SVM termed as GA-SVM technique to appropriately choose feature subset. The redundant feature reduction of the anticipated approach enhances classification performance is evaluated with five diverse disease dataset.

Jin et al., (2014) offered a wrapper technique for CKD prediction using the following steps: 1) model is produced from data mining; 2) attribute evaluation with wrapper sub-set and finest search methodology are used to choose features; and 3) classification algorithms are used. The outcomes obtained from the experiment is enhanced for reducing dataset measured with original dataset. A well-constructed model is used for improving CKD quality. This structure comprises of three processes such as ensemble learning, classification and ensemble learning. k-nearest neighbor (kNN) and Correlation-based FS (CFS) classifier outcomes in superior classification accuracy. The author modeled CKD identification techniques are used for filter and wrapper approaches. The results depict the reduction in number of features does not fulfill effectual performance. Table 2.2 depicts various ML based algorithms which aids in predicting the disease.

Table 2.2 Various ML algorithms used for disease prediction

S.No	Predicted disease	Algorithms	Data types	Cross-validation (CV)	Best algorithm
1	Kidney disease	ANN, K-NN, DT, NB	manual and clinical based data measurement	Validation	DT
2	Liver disease	LR, RF, ANN, SVM	Ultra-sonography test data,	10-fold CV	LR

			demographic data		
3	Lung cancer	SVM, RF, DT	Demographic and clinical data	10-fold CV	RF
4	Micro RNA	SVM, RF	microRNA data	10-fold CV	RF
5	Parkinson's disease	SVM, ANN	Demographic and voice recording	5-fold CV	SVM
6	Parkinson's disease	SVM, k-NN	Demographic and voice recording	10-fold CV	k-NN
7	Parkinson's disease	SVM, NB, k-NN	Demographic and voice recording	10-fold CV	SVM
8	Prostate cancer	SVM, NB, DT	MRI data	Leave-one-out	SVM
9	Prostate cancer	NB, DT	Clinical data	10-fold CV	NB
10	stroke	SVM, LR, ANN	Demographic data and electronic medical claim	10-fold CV	ANN
11	Diabetes	SVM, LR, ANN	Electro-chemical measurements	3-fold CV	SVM
12	Diabetes	SVM, LR, k-NN	Manual and clinical laboratory report, anthropometric	5-fold CV	SVM

			and demographic results		
13	Diabetes	SVM, RF, NB, LR, k-NN	Clinical and demographical results	5-fold CV	RF
14	Diabetes	SV, RF, LR, SVM	Manual and clinical laboratory report, anthropometric and demographic results	10-fold CV	SVM
15	Diabetes	DT, NB, SVM	Clinical data	10-fold CV	NB
16	Diabetes	SVM,RF	Gene expression and clinical data	10-fold CV	RF
17	Heart disease	SVM, RF, k-NN	Signal data	---	RF
18	Heart disease	SVM, NB, ANN	Demographic, clinical and image data	---	NB
19	Heart disease	NB, DT, ANN	Demographic and clinical data	2-fold CV	NB
20	Heart disease	RF, LR	Electronic health records	5-fold CV	LR

2.9. Reviews on classification and prediction techniques

There are number of investigations are achieved to diagnose and predict CKD using ML approaches. These approaches are used for prediction and classification in bio-medical fields.

2.9.1. k-Nearest Neighbor

k-NN is utilized by various investigations for handling problems related to classification. Hashi et al., (2017) utilized k-NN for CKD over dataset from Pima Indians data. The outcomes demonstrate k-NN gives 76% accuracy with minimal error rate using CKD prediction. The author develops a medical DSS with multi-layer classifier for disease prediction. The author used k-NN for prior disease prediction. With the application of k-NN, the author used 57% accuracy of Statlong dataset and 59% accuracy over Cleveland dataset. The author uses k-NN for liver disease prediction over dataset considered. K-NN outcomes demonstrate 62% accuracy rate and 0.3719 error rate respectively. Fig 2.2 depicts the pictorial representation of k-NN.

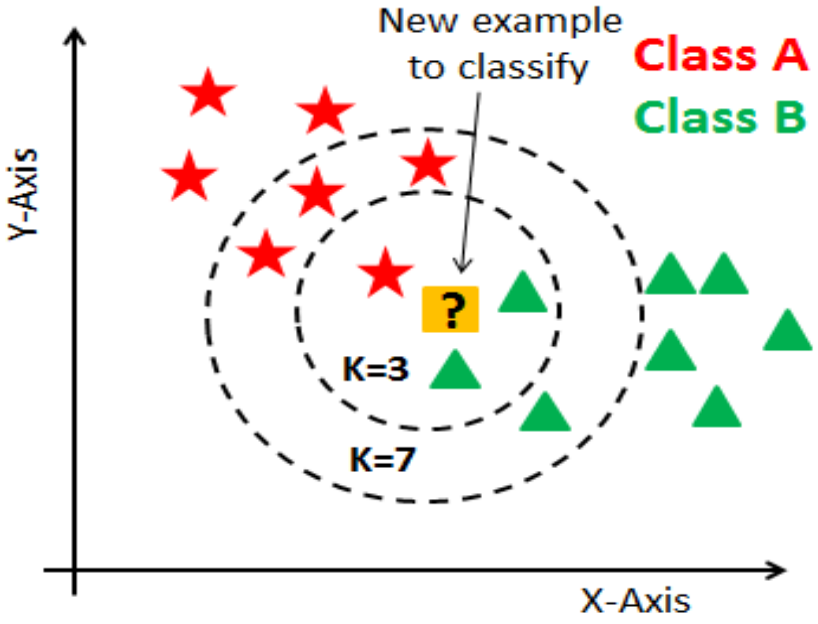

Fig 2.2 k-NN classifier

2.9.2. Decision Tree

Basically, DT is utilized for classification issues with the competency to generate optimal outcomes. Quinlan et al., (1986) used DT for kidney disease prediction. The author performed experimentation on dataset considered. With the use of 24 features from dataset, 98% accuracy and 0.0160 error rate is attained, when using eight features, the author attains 98% accuracy and 0.0160 error rate respectively. Azmeen et al., (2013) applies this approaches for dengue fever prediction from diverse hospitalized patients. The outcomes attained from DT are 76% accuracy to categorize data as negative or positive. Archarya et al., (2015) uses DT to predict the fatty liver disease over ultra-sound images. The author adopts hundreds of ultrasound images in this work. The results attained from DT outperform other approaches for producing 99% accuracy. Fig 2.3 depicts the pictorial representation of DT.

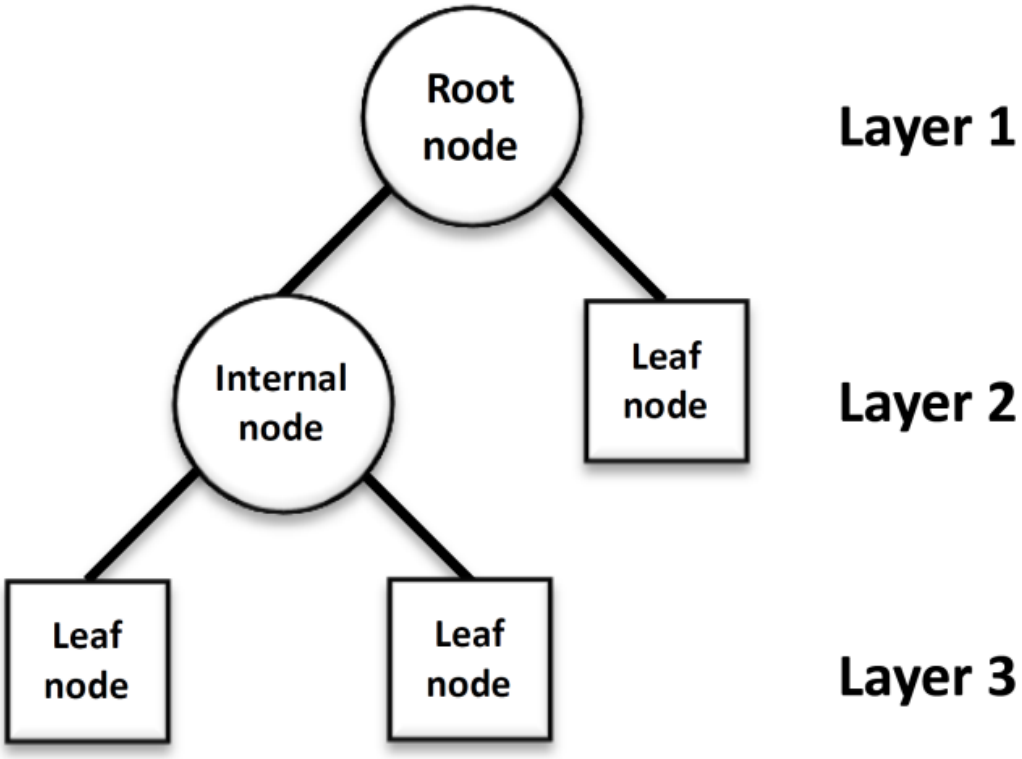

Fig 2.3 Basic structure of Decision Tree (DT)

2.9.3. Artificial Neural Network (ANN)

ANN is extremely popular and extensively applied in a technique that uses diverse classification issues. Baitharu et al., (2017) applied these techniques for healthcare DSS with liver disease dataset. The author used this system for accuracy measurement and error rate prediction. The outcomes provide superior accuracy for ANN evaluation with 71% and nominal error rate of 0.3543.ANN is used by Vijayarani et al., (2015) for kidney disease prediction over a dataset attained from diverse laboratories, and hospitals. Based on the attained outcomes, ANN classification accuracy is 87%. The author uses particle swarm optimization and artificial bee colony for optimizing neural network in heat prediction and cooling issues of residential apartments. With ANN on trained data, error rate attained is 2.5730 and testing data is 2.4280 respectively. Fig 2.4 depicts the pictorial representation of ANN.

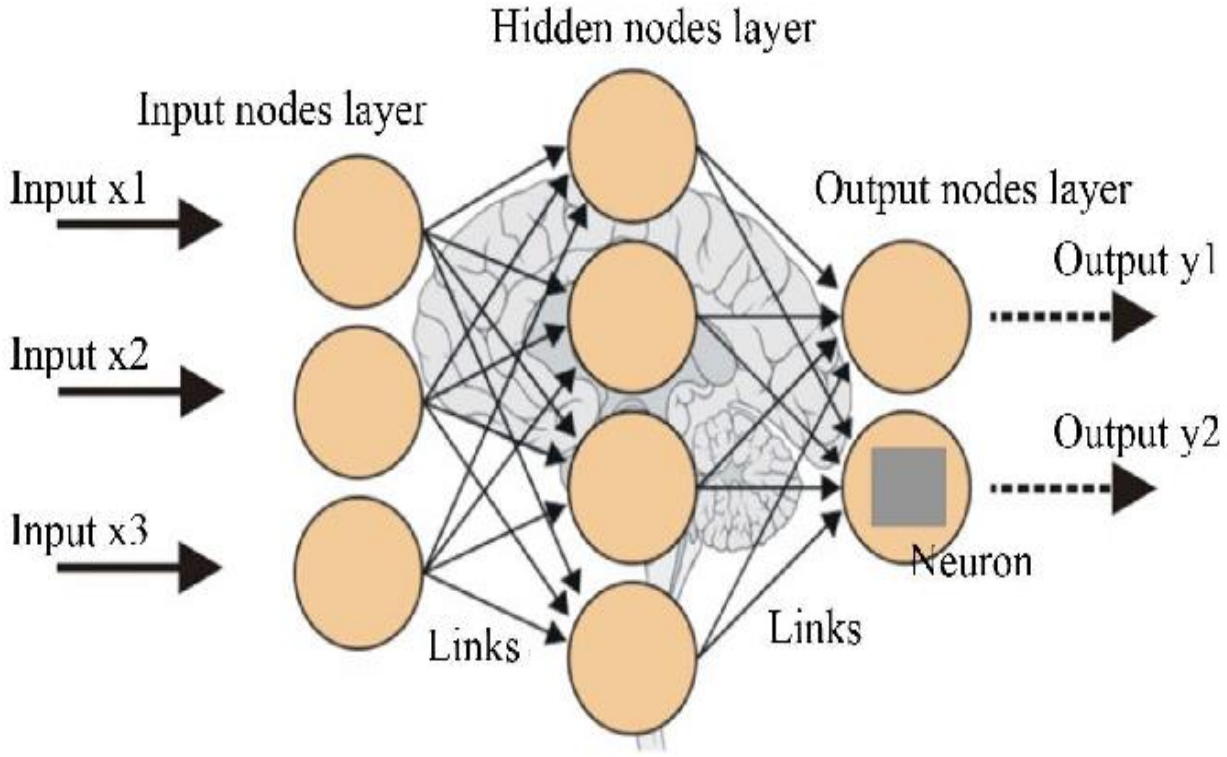

Fig 2.4 Artificial Neural Networks (ANN)

2.9.4. Probabilistic Neural Network (PNN)

PNN is a FFNN that comprises of three layers known as input, hidden and output layers. PNN is used by Archarya et al., (2016) to model a DSS for predicting fatty liver disease with ultrasound images. The author performed this experimentation of data from Malaysia. The study suggests that PNN as finest classifier reporting accuracy of 98%. The author used PNN for kidney disease stages prediction. As an outcome, the author recommended PNN outperforms other methods with an improved accuracy of 96% respectively. Fig 2.5 depicts the pictorial representation of PNN.

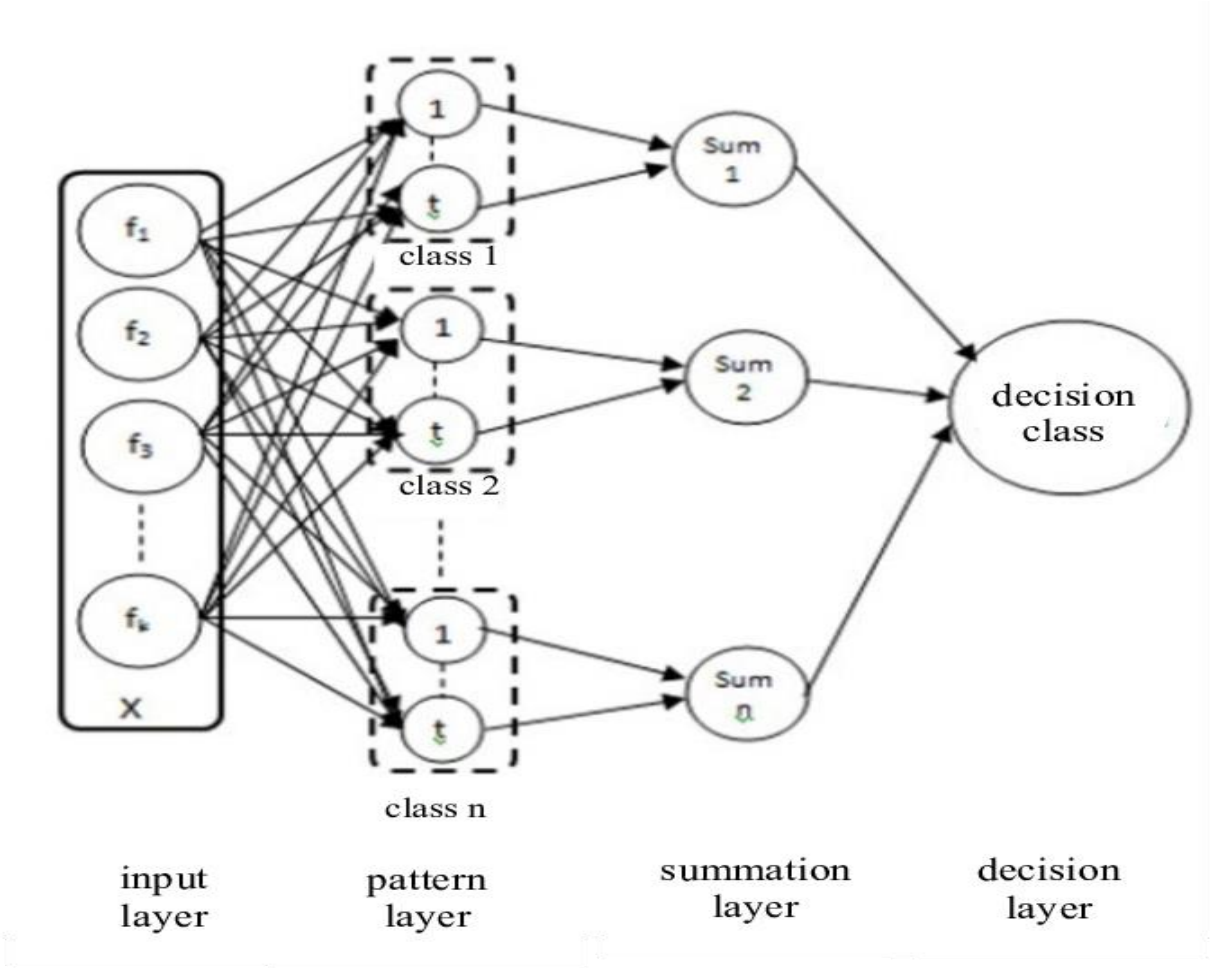

Fig 2.5 Probabilistic Neural Network (PNN)

2.9.5. Naïve Bayes

NB is a classification approach which is based on probability and are utilized for diverse disease prediction. Rish et al., (2001) deployed NB in predicting the kidney-related disease over dataset considered which is comprised of 24 attributes and 400 populations. With these population of 150 are NOT CKD and remaining of 250 are prior phase 250. With the 24 attributes, the author attained 99.5% accuracy and 0.0060 of error rate. NB is used by Ameer et al., (---)for profiling gender and age with diverse feature types. The author concluded that the outcomes for gender prediction over hotel reviews applies POS features where NB based accuracy is 60%, and Sn-gram (POS) features based NB accuracy is 51%. Fig 2.6 depicts the pictorial representation of NB.

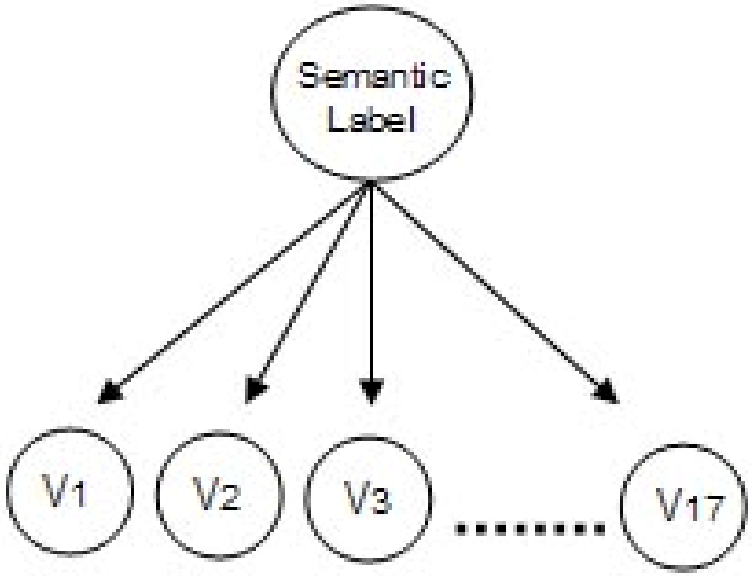

Fig 2.6 Simple Bayesian network structure

2.9.6. Random Forest

RF is a cooperative learning model that uses classification and regression issues. Manish et al., (2016) utilized RF for predicting liver disease with dataset considered. The outcomes of RF show finest accuracy with 73% and 0.3804 error rates. The author utilized RF for revealing temporal pattern for dengue incidence with Meterological factors like Philippines, metropolitan

Manila metropolitan respectively. The author used two diverse datasets for experimentation, Meteorological Factors and successive Lagged MF respectively. With this experimentation, RF error rate is 0.24 over Meteorological dataset and 0.16 over Lagged MF dataset. Fig 2.7 depicts the pictorial representation of RF.

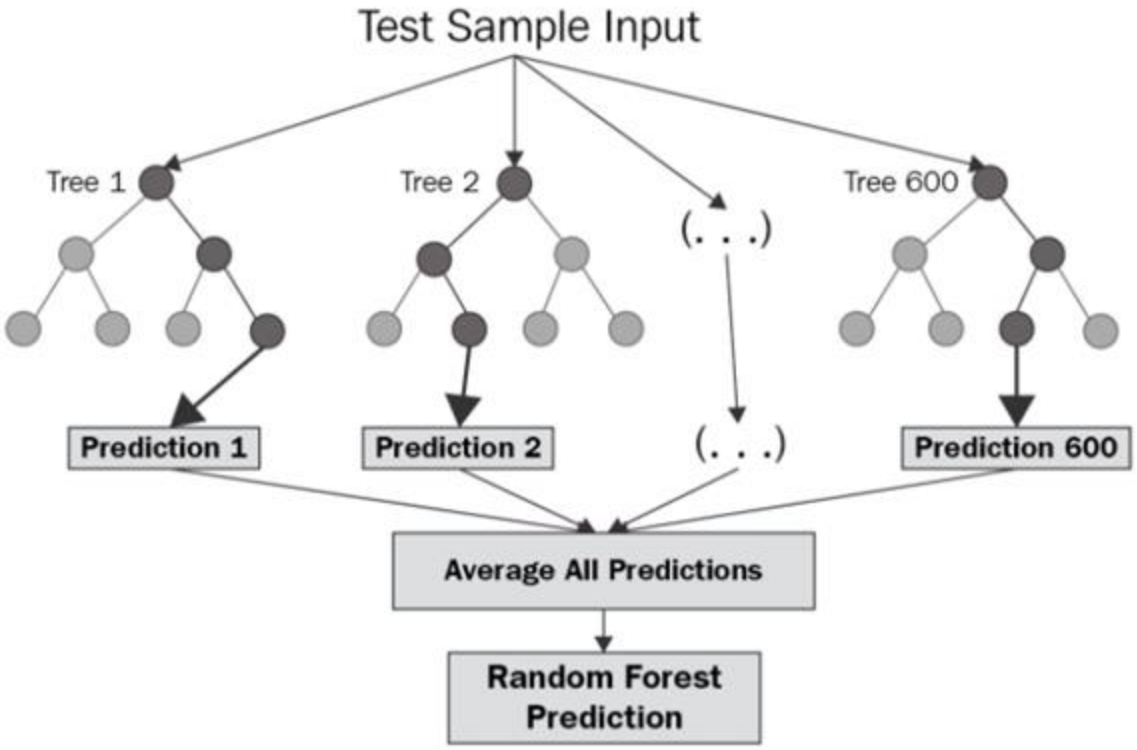

Fig 2.7 Random Forest model

2.9.7. AdaBoost (AB)

Basically, adaboosting is the arrangement of probable technique for developing dynamical approaches. Acharya et al., (2016) utilized DSS with GIST descriptors for extracting ultrasound images. The outcome uses AB with an accuracy of 94% and 92% sensitivity respectively. The author used AB for prediction of fatty liver diseases with discrete cosine transform and random transform features over ultra-sound images. AB is used by Yip et al., (2016) to model the parameter-based on fatty liver which is non-alcoholic with the population. Fig 2.8 depicts the pictorial representation of Adaboosting.

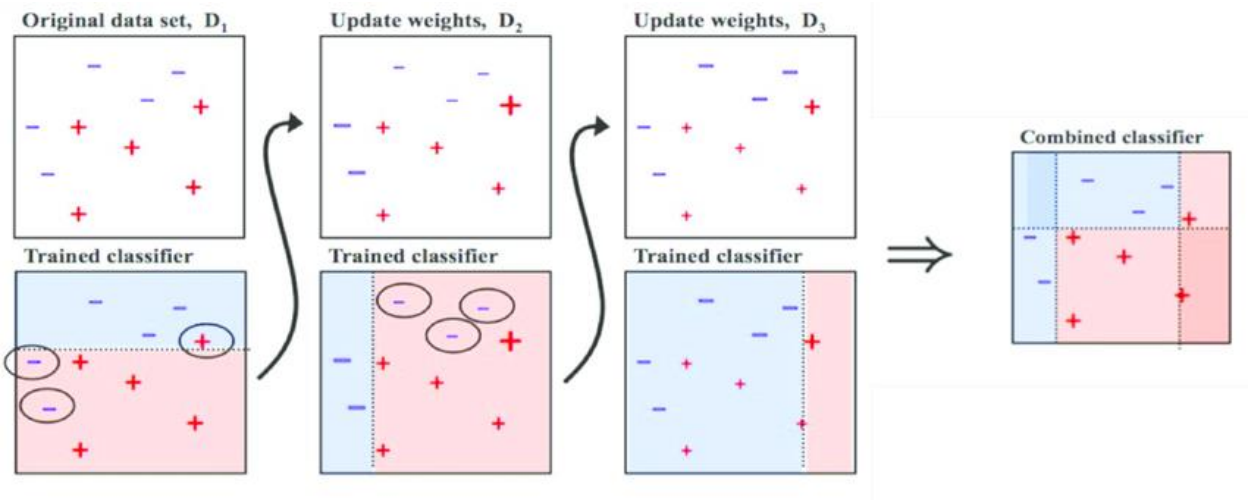

Fig 2.8 Adaboosting

2.9.8. Support Vector Machine (SVM)

SVM is considered as a most influencing learning approach that is based on the present enhancements over the statistical theories used for learning. Jyothi et al., (2015) used SVM for classifying the liver based patient data over the UCI ML repository. With the original dataset, the author attained 71% accuracy while performing sampling the author attained 68% accuracy. Fig 2.9 depicts the pictorial representation of SVM.

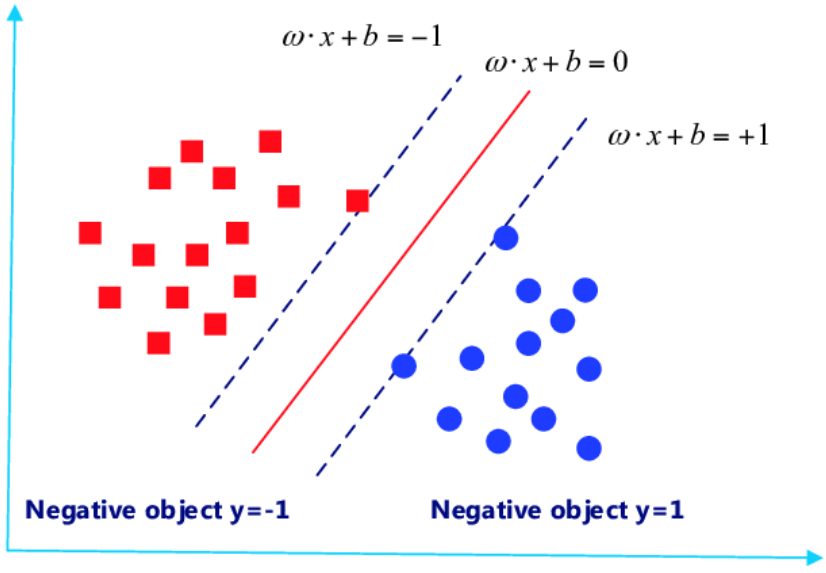

Fig 2.9 Support Vector Machine (SVM)

The author used SVM for classifying model permeability and lithofacies over heterogeneous reservoirs. With SVM, clustering based on training error rate is 1.1497 respectively, while during core training error rate is 0.6377, core clustering based permeability prediction error is 1.5179, log clustering permeability error rate is 1.2885 respectively.

2.9.9. Logistic Regression (LR)

LR quantities are considered as an association among least one independent variable, and continuous dependent variables continuously from the most part that uses likelihood scores as prediction values of dependent variable. The variations are considered as the proportion among the success probability over failure probability, that is, $p_i = (1 < p_i)$, where $'p'$ is probability of model with class $'0'$. In some conditions, when $p > 0:5$; the for an instance, the value is provided as class 0. However, it is given to make a decision with clas 1. As , the computed output probability is based on various condition – the coefficient refers to all p_i. The proportion variations are based on exponential weights. The coefficients are weighted certainties that are used for every attribute before considering them together. In some cases, the results are based on probability newer occasion has to be placed with class yes (> 0.5) respectively. Fig 2.10 depicts the pictorial representation of LR.

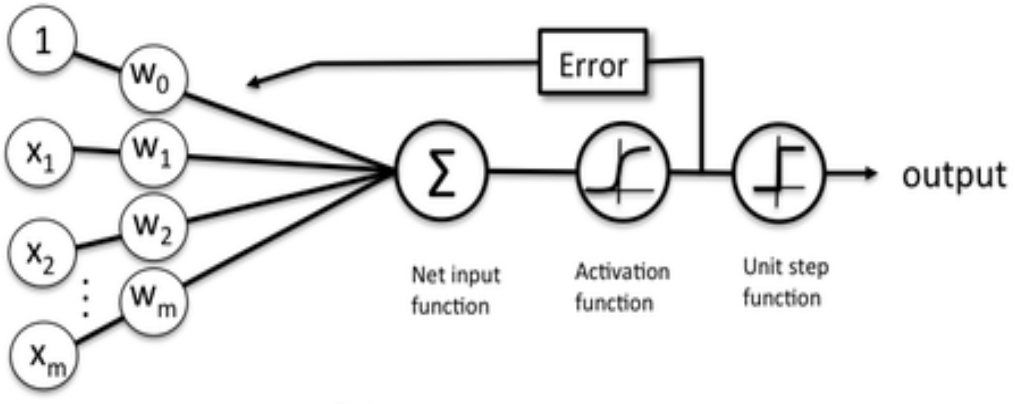

Fig 2.10 Logical regression

2.9.10. Multi-Layer Perceptron (MLP)

MLP is considered as the most significant classes over neural networks that includes input layer, output layer, and minimum single hidden layer respectively. This is used viably over various applications to deal with various and troublesome issues by employing them in a supervised manner using well-known approaches, that is, back-proliferation approaches (Vijayarani et al., 2015). These techniques are based on error-rectification learning instructions. With all these, it is considered as a speculation of diverse versatile filtering model. Fig 2.11 depicts the pictorial representation of MLP.

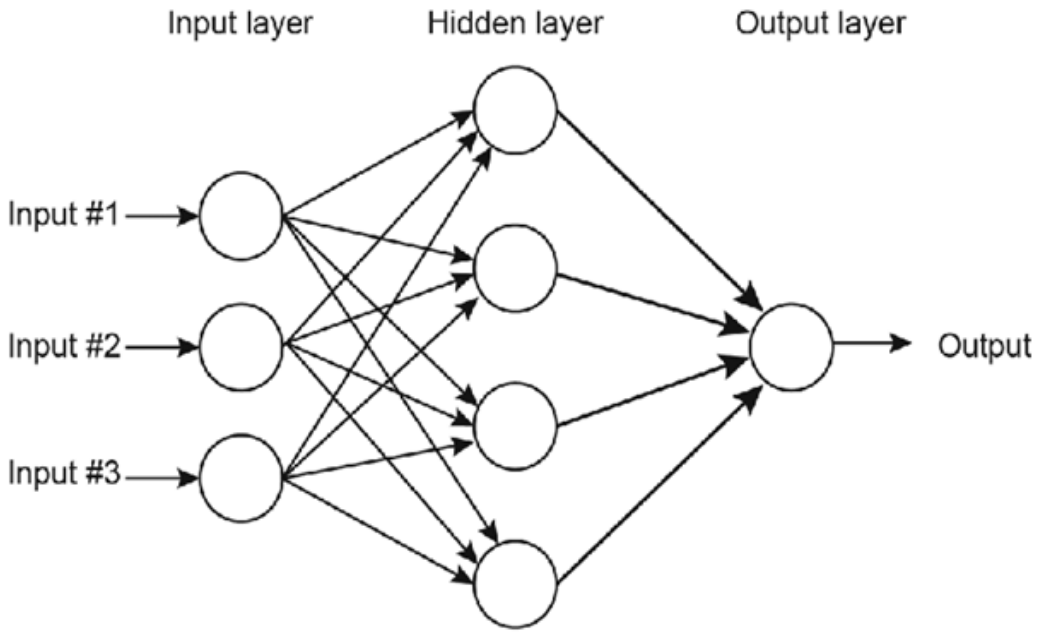

Fig 2.11 Multi-layer perceptron

2.9.11. J48 Decision Tree

J48 is measured as an advanced C4.5 rendition. This methodology is applied for dividing and conquering schemes. It is used as a pruning scheme to model the tree (Quinlan et al., 1986). Typically, this technique is considered as the entropy measure and information gain. Therefore, it provides tree structure with root, leaf and moderate nodes respectively. These nodes holds gain

and decision based outcomes. Fig 2.12 depicts the pictorial representation of DT. Table 2.3 depicts the advantages and disadvantages of ML algorithms.

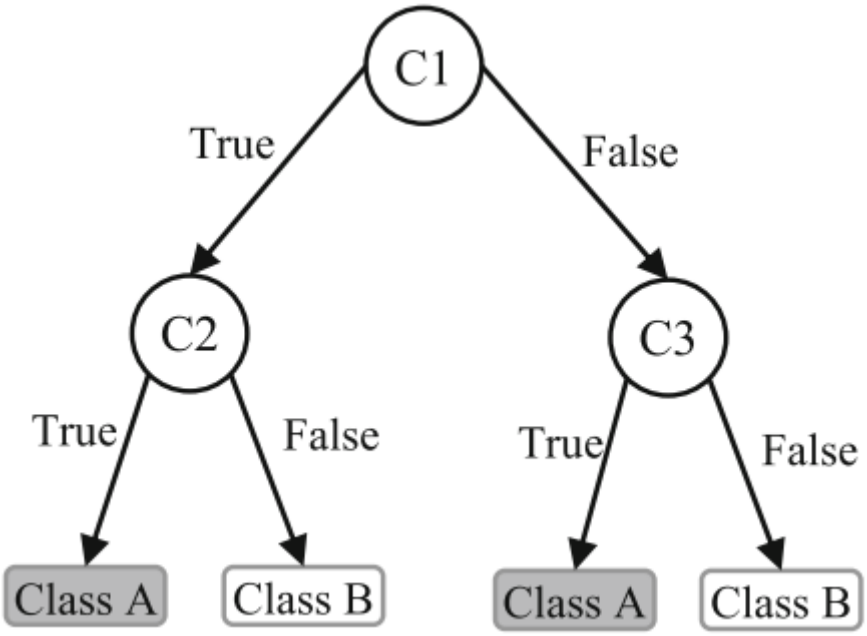

Fig 2.12 Sample Decision Tree

Table 2.3 Various ML algorithms with its advantages and disadvantages

S.No	Algorithm	Advantages	Disadvantage
1	**ANN**	- predict complex non-linear relationship among independent and dependent variables - available training algorithm - perform both classification and regression	- possess black box characteristics - does not possess exact decision-making process - complex classification problem - computationally expensive
2	**DT**	- classifier tree is easier to interpret and understand	- classes needs mutual exclusion

		- easier data-processing - support multiple data types - generate robust classifiers	- cannot bran attribute values when non-leaf nodes are missing - depends on attribute variables - lesser performance
3	**k-NN**	- simpler algorithm and instance classification - handle noisy instances with missing attribute values - perform both regression and classification	- computationally expensive during the process of higher amount of attributes - attributes are provided with higher significance which leads to poor classification
4	**LR**	- simpler and easier to implement - model can be updated easily - needs lesser amount of data - easier probabilistic interpretation of model parameters	- does not possess better accuracy when input shows complex relationship - no proper linear relationship among variables - key components are vulnerable - over-stated prediction accuracy based on sampling bias
5	**NB**	- simple and useful for handling larger dataset	- classes are exclusive mutually

		- can deal with both multi-class and binary classification problems - needs lesser amount of training data - performs probabilistic prediction and deal with both discrete and continuous data	- presence of dependency among negative attributes after classification - considers normal distribution as numeric attributes
6	**RF**	- lesser chance of over-fitting and variance - ensemble classifier performs better classification than individual classifier - scales larger datasets - perform estimation with attributes and variables	- computationally expensive and complex - number of base classifiers has to be defined - attributes take higher number of values for estimating variable significance - over-fitting is easier
7	**SVM**	- more robust - handles multiple feature space - reduced risk for over-fitting - works well during classification of semi-structured/unstructured data	- computationally expensive for complex and large dataset - lesser performance over noisy data - weight, resultant model, and variables are complex to predict - SVM cannot categorize two classes

2.10. Current Approaches to Medical Decision Support Systems

An approach known as C4.5 Rule-PANE integrates the benefits of ANN ensemble with rule induction as stated by Vijayarani et al., (2015). In general, the ensemble ANN is followed, trained by the generation of training dataset by providing feature vectors of original training instances used as instances which are trained. Manipulate the class labels of original training instances with the ensemble output. Further, the training data is concatenated with random generation of feature vectors and measure the same with corresponding class labels output. However, rules from newer training data sets are learnt with rule induction approaches termed as C4.5 rule. Some case studies with hepatitis, diabetes, and breast cancer depicts that C4.5 Rule-PANE is competent of producing rule with superior generalization capacity that advantages from ANN ensemble indeed of stronger lucidity, which advantages in rule induction.

Lambodar et al., (2011) anticipated a data mining model for biological datasets. This is applied over the Hepatitis B Virus DNA real-world dataset. The superior predictive accuracy of the classifier is proven with the experimental outcomes. This is more advantageous with the prediction of liver cancer. Moreover, the author presents the classification approach that functions over the basis of non-linear integral. The fuzzy measure is used and non-linear integral over the system is provided with effectual results based on the fact that non-additivity of fuzzy measure shows the importance of feature characteristics along with inherent interactions.

The susceptibility of liver cancer prediction, chronic hepatitis from single nucleotide polymorphism (SNP) data with the use of Machine Learning approaches, decision tree, SVM, and decision rules are provided. The author analyzed a set of SNPs which is suitable for the disease as well. Moreover, it is used for backtracking approaches to merge the feature selection process, backward elimination, and forward selection and demonstrates the technique with huge benefits in determining finest outcomes by the experimentation. After a widen experimentation evaluation, it is known that, the decision rules are competent of differentiating chronic hepatitis from normal with maximal accuracy of 73%, whilst SVM is 67%, and decision tree with an accuracy of 72% respectively. However, it is demonstrated with decision rules and decision tree has potential tools for diagnosing susceptibility to chronic hepatitis from SNP data.

Three diver NN algorithms are anticipated by Yildirim et al., () for the prediction of hepatitis diseases. The disease comparisons are attained with lesser statistical techniques used in the earlier stages. As, the classification accuracy of MLP is lesser, it is known to be considered as a finest option. However, constant performance for single run is based on random weighted initialization in training was not fulfilled. Nonetheless, CS-FNN possesses superior classification accuracy during hepatitis diagnosis. The outcome demonstrates that the utilization of hybrid network CS-FNN integrates RBF and MLP was extremely constant for prediction. The comparison outcome demonstrates that NNs are beneficial equivalently for the hepatitis disease prediction towards traditional statistical approaches.

Integer-coded genetic algorithm (GA) and Support Vector Machine (SVM) are utilized for heart disease classification as depicted by (Jyothi et al., 2015). The simpler SVM algorithm is utilized for determining support vectors in an iterative, faster manner. An integer-coded genetic algorithm is utilized to demonstrate the significance of relevant and appropriate features and to eliminate the redundant and irrelevant features. Cleveland heart disease database is utilized and comprises of 303 cases partitioned into 5 classes, everything with 13 diagnostic features. The author noticed 5-class classification problem to specify the increase in overall accuracy of 72%, while utilizing optimal feature subset and two class problem with an accuracy of 90%.

Vijiayarani et al., (2015) anticipated multi-layer perceptron based DSS to assist heart disease diagnosis. Here, a total of 352 medical records are utilized to train and test the system. Three evaluation methods like hold-out, cross-validation and bootstrapping are used to evaluate system generalization for classifying fiver diverse kinds of heart diseases. The anticipated MLP-based DSS can attain diagnostic accuracy of (<90%).

2.11. Reviews on Assessment criteria

Some assessment criteria like MAE, RMSE, RAE and RRSE are used to evaluate classification of error rate approaches, precision; F-measure, recall and accuracy are utilized to compute the

performance accuracy of techniques used. All these evaluation are done with the following conditions.

2.11.1. Mean Absolute Error (MAE)

Generally, MAE is an average of absolute errors that are evaluated with Eq. (2.1):

$$MAE = \frac{1}{2}\sum_{j=1}^{n} |y_i - y| \tag{2.1}$$

Here, $'n'$ is number of errors, $|y_i - y|$ is absolute error.

2.11.2. Root Mean Squared Error (RMSE)

It is a quadratic scoring rule that computes average error magnitude and expressed in Eq. (2.2):

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_i - 1)^2} \tag{2.2}$$

2.11.3. Relative Absolute Error (RAE)

RAE is compared with simpler predictor which is known as an average of real values. However, the error is the total absolute error than complete squared error. It is expressed in Eq. (2.3):

$$RAE = \frac{\sum_{j=1}^{n} |p_{ij} - T_j|}{\sum_{j=1}^{n} |T_j - T|} \tag{2.3}$$

Here, p_{ij} is value predicted by certain model $'i'$ for record $'j'$ (out of records), T_{ij} is target value of records $'ij'$ and TN is considered as perfect fit with numerator which is equal to 0.

2.11.4. Root Relative Squared Error (RRSE)

It is an average of real values. However, relative squared error revenues out of squared error and standardizes partitioning the complete squared error of modest, predictor. With square root of relative squared error reduces the error to similar dimensions as quantity prediction.

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{ij} - T_j\right)^2}{\sum_{j=1}^{n}\left(T_j - T\right)^2}} \tag{2.4}$$

2.11.5. Accuracy

Accuracy is an instinctive performance measure and ratio of appropriately predicted or classified perceptions to entire perceptions. It is expressed as in Eq. (2.5):

$$Accuracy = TP + \frac{TN}{TP} + FP + FN + TN \tag{2.5}$$

Here, TP is number of positive classifications, FN is number of false negative classification, TN is number of true-negative classifications, and $FP$is number of false positive classifications.

2.11.6. Precision

Precision is a ratio of positive observations effectually to all-out positive observations. It is expressed as in Eq. (2.6):

$$Precision = \frac{TP}{TP} + FP \tag{2.6}$$

2.11.7. Recall (Sensitivity)

It is depicted as the ratio of appropriately positive observations to all observation in appropriate classification. It is expressed as in Eq. (2.7):

$$Recall = \frac{TP}{TP} + \text{FN} \tag{2.7}$$

2.11.8. F-Measure

F-measure is a weighted average of recall and precision. The scores are considered as both false positives and negatives. Specifically, it is a straight forwards precision, but, F1-measure is extremely essential than accuracy; specifically, it has lop-sided class distribution.

$$F - measure = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{2.8}$$

In some cases, 10-fold CV is used for all classifiers to compute the performance. This is a process that partitions the dataset into 10 subsets of equivalent sizes; In which, one subset is used for testing whereas others are used for training. It is performed until each subset is used for testing. Standard scheme for evaluation is 10-fold CV.

2.11.9. Confusion Matrix

To evaluate the anticipated model performance, the evaluation metrics considered for this research are confusion matrix which is represented in Table 2.4.

Table 2.4 Confusion Matrix

	Number of Records	**True Outcome: Patients having CKD disease)**	
		P (Patients having CKD)	**N (Patients who do not have CKD)**
Actual Class	**P (Patients having CKD)**	**TP** (Patients who are correctly diagnosed as CKD	**FP** (Patients who are having CKD but wrongly diagnosed as normal)
	N (Patients do not have CKD)	**FN** (Patients who are normal, but diagnosed as CKD)	**TN** (Patients who are normal and diagnosed as normal)

Where, TP- number of True Positive (RA) is accurately classified as CKD. TN – number of True Negative (Non-CKD) is inappropriately classified as Non-CKD. FP- number of False Positive (i.e. non-CKD is classified as CKD), FN-number of False Negative (CKD is classified as non-CKD).

2.12. Research gaps

Even with the above mentioned process, there is some research gaps found with the prediction of CKD in earlier stage using Machine Learning approaches. Those research gaps are listed below:

1. There are no proper investigations for predicting CKD with Machine Learning approaches in India regarding the rate of eGFR decline.
2. No proper studies in India regarding the reduction of CKD progression.
3. No proper studies are done in India for randomized controlled trail based on CKD progression.

2.13. Summary

The occurrence of CKD is increasing and the available studies discussed that hyper-tension and diabetes are the common cause of CKD over the subjects in India. CKD prediction is earlier stage is extremely significant to delay the disease progression that reduces the economic burden of families, individuals and communities. More such investigations are needed to sensitize people based on kidney functionality. The literature review helps in researchers to have a deeper insight towards the progression of present studies. There are various research inputs in this field. With all these literatures, there are some drawbacks that need to be identified with well-modeled predictor design specifically for randomly controlled trails. The reviews assist in predicting the research gaps; construct the implementation, study design and proposal for further data analysis.

CHAPTER 3

PREDICTION OF CKD RISKS USING MACHINE LEARNING APPROACHEs

3.1. Prologue

The target objective of this chapter is to model a prediction framework for chronic kidney disease (CKD) prediction using learning approaches. There are some prevailing approaches that are used for examining the CKD disease severity from the physical, clinical and laboratory examinations. The evolution of data mining approaches paves the way for CKD prediction through learning approaches from the case histories. This research work concentrates on using on NB and CbH to provide a suitable precision measure with hierarchical classification. The evaluation metrics gives better results and reliability towards the predictor model. The classification approaches is chosen based on an extensive study with existing approaches in modeling a scheme for medical applications.

3.2. Research objectives

The objective of this prediction model is:

- ✓ To design a prediction model for identifying the influencing factors related to CKD prediction.
- ✓ To gather the clinical records from the diabetes analysis centre globally.
- ✓ To model a classification approach into CKD or Non-CKD groups based on analyzing the features.
- ✓ To attain better precision outcomes from choice based hierarchical categorization.
- ✓ To perform appropriate comparison with the prevailing approaches to promote the significance of the anticipated model.

These are the objectives associated with the present study for predicting the occurrence of CKD with effectual assessments.

3.3. Research Methodology

The objective is to compute the earlier prediction of CKD termed as severe renal failure with ML approaches with recommendation of selecting hierarchies to attain appropriate results with

precision by computing performance on sensitivity. The learning scheme characteristics are based on information collection with stabilized outcomes over developed model. Knowledge acquisition from the repositories is known as information extraction. Indeed of learning from clinical base heart ailment from 600 clinical records are attained from diabetic analysis centre. Information set is validated from classification based on NB and CbH. With classification evaluation in NB and choice based hierarchies, the experimental results gives better precision with 90% for the anticipated model. Fig 3.1 depicts the flow diagram of prototype model.

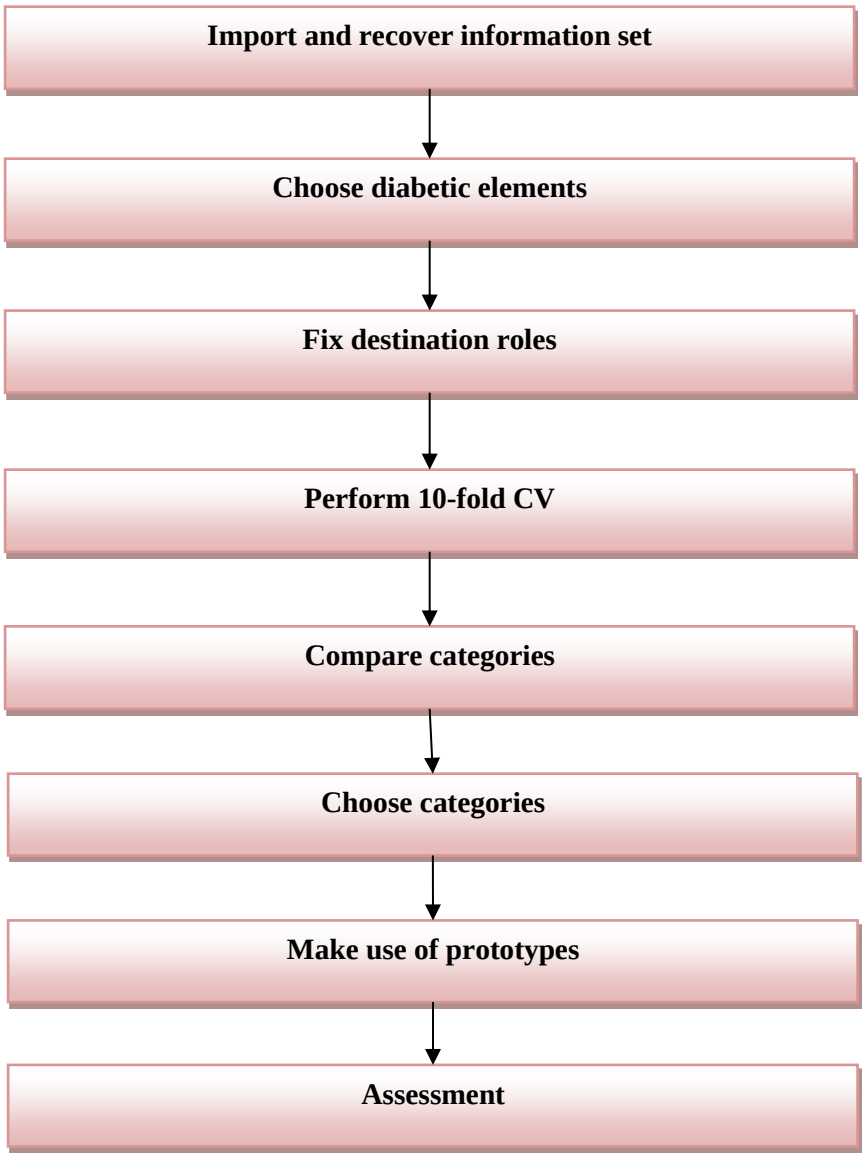

Fig 3.1 Proposed Prototypes

3.3.1. Techniques

The feature or information extractions are used by various investigators for predicting the disease severity. The predictions of these diseases are carried out with the combination of evaluating machine learning, medical repository and statistical practice. The block diagram of anticipated NB classifier is given in Fig 3.2.

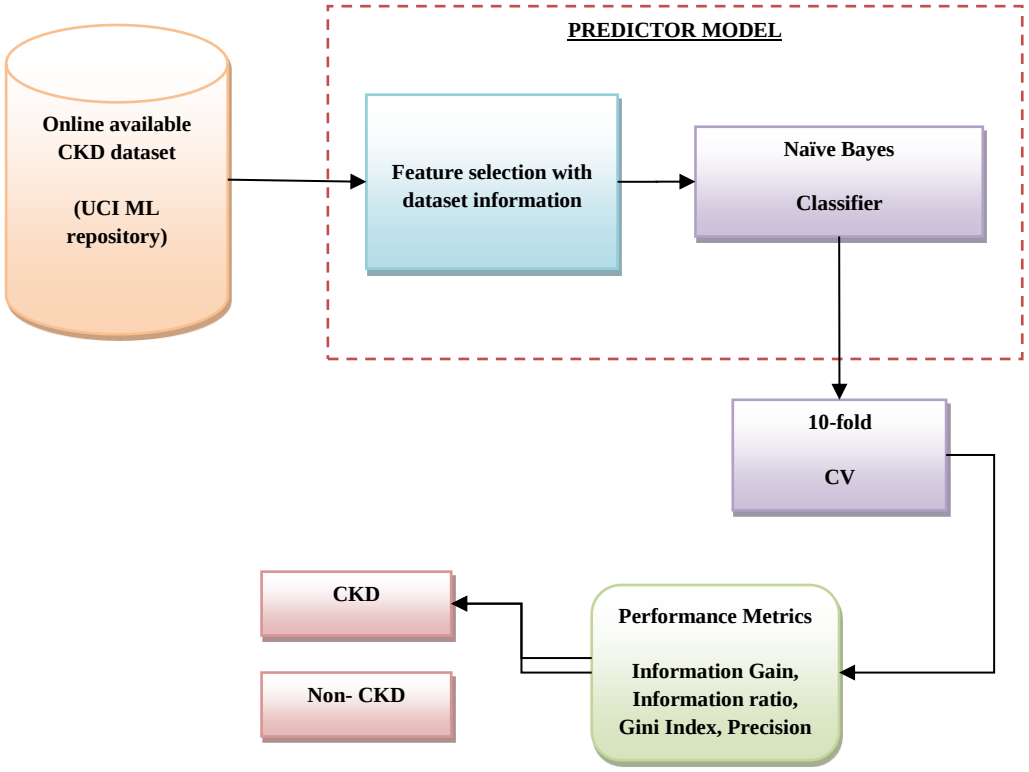

Fig 3.2 Block diagram of proposed NB predictor model

The existing approaches are based on classification and evaluation that includes feature extraction scheme. It is included with the construction of information for classified analytical prototype, instructions, and validating the classification efficiency. The evaluation of these activities are attained and based on classification information. Various sections are used for the information extraction using software for validation purpose that results in effectual authorization. This is used for effortless validation of prototypes that assists in action synchronization with text and information extraction, statistical and business analytics. Information extractions are based on

YALE and WEKA which makes JAVA utilization with abundance opportunities over the knowledge base, business, scientific, and medical researchers.

The faster information extraction helps in designing a 10-fold cross-validation of ML approaches for enhancing performance. The validation is performed in data with 90:10 ratios for testing and training of the provided data. Cross-validation helps in the development of set of information for set of information development for instructions. The development is performed with the anticipated model for error rate computation to overcome the disadvantages in prevailing models. Element-based information extraction of every record is given in Table 3.1.

Table 3.1 Elements Used for Assessment

	Elements	**Variety**	**Details**
Risk Factor	Gender	Binomial	Female/Male
	Age	Integer	Patient's Age
	Genetics	Polynomial	Father/Mother
	Weight	Number	Patient's Weight
	Smoking	Number	Yes/No
Investigation	BP	Polynomial	Patient's BP
	Fasting	Integer	Fasting Blood Sugar
	PP	Integer	Post Prondial Blood Glucose
	AIC	Number	Glycosylated Haemoglobin Test
	LDL	Integer	Low-Density Lipoprotein
	VLDL	Integer	Very Low-Density Lipoprotein
	HDL	Number	High-Density Lipoprotein
	Urea	Number	Urea Creatine
	Threat Category	Polynomial	Threats of Kidney Disease

3.3.2. Naïve Bayes classifier

It executes probabilistic classifier is employed in this work. It is extremely simple on diverse kinds of datasets. Also, it provides reliable and good results. This classifier has the ability

to achieve higher degree of prediction accuracy in an extensive variety of applied contexts. Assume, $'X'$ is a data with no class labels. Consider, $'H'$ is hypothesis where $'X'$ belongs to certain class $'C'$. With work pretends to establish, where the probability that hypothesis $'H'$ is related to observed $'X'.P(H|X)$ is posterior probability specifying confidence in hypothesis where $'X'$ is provided. Bayesian theorem gives posterior probability $P(H|X)$ computation using probability. The preliminary association with Bayes is specified as $P(H|X) = \frac{P(X|H).P(H)}{P(X)}$. Fig 3.3 depicts the NB classifier model. Similarly, the flow diagram of anticipated model is given in Fig 3.4.

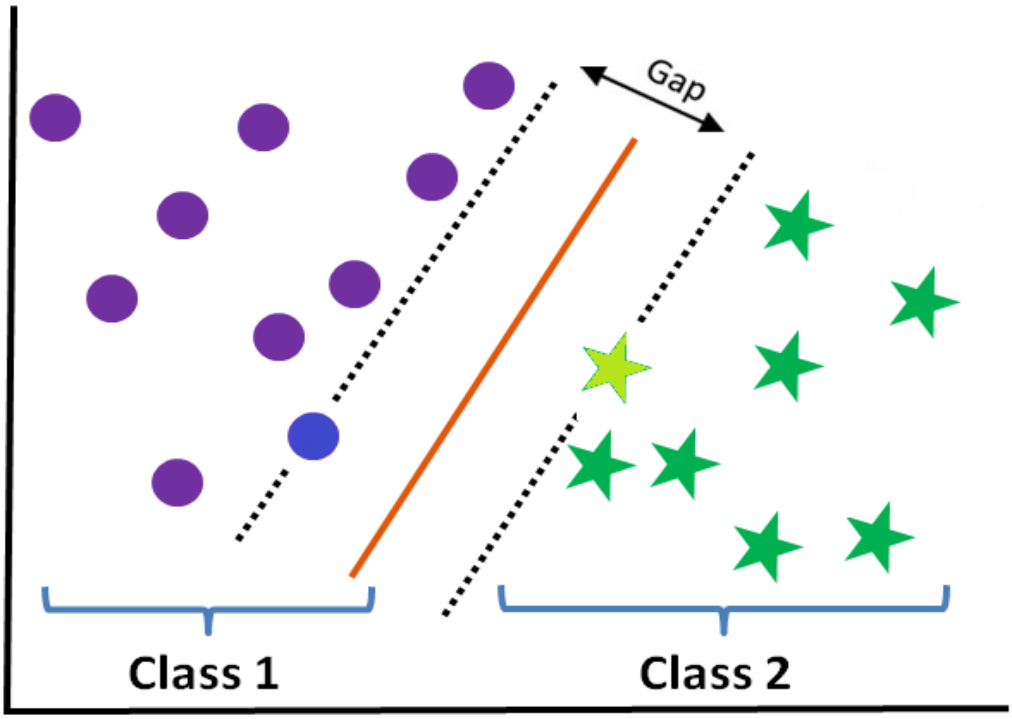

Fig 3.3 NB classifier model

When there are set of $'n'$ samples, $S = \{S_1, S_2, \dots, S_n\}$ where samples S_i is specified as a n-dimensional feature vectors $(X_1, X_2, .., X_3)$. This is considered as training dataset with X_i which corresponds to $V_1, V_2, \dots, V_m$features respectively. As well, there are $'n'$ classes where $c_1, c_2, .., c_n$ and every other samples belongs to one of these mentioned classes. In this model, the value of $'n'$ is two where these two classes prevails. An extra data samples are provided where the class are not identified, it is probable to identify the classes for those samples with higher conditional probability $P(C_k|X)$, where $i = 1,2, \dots, n$. The preliminary idea for NB is expressed as in Eq. (3.1):

$$p(C_k|X) = \frac{P(X|C_k).P(C_k)}{P(X)} \quad (3.1)$$

Here, $P(X)$ is constant, where the product is based on $P(X|C_k).P(C_k)$ should be maximal. The prior probabilities of these classes are evaluated with Eq. (3.2):

$$P(C_k) = \frac{Numberoftraininginstacnesofclass C_k}{n\ ('n'istotalnumberoftraininginstances)} \quad (3.2)$$

With the conditional independence assumption among the attributes, the values are expressed as in Eq. (3.3):

$$P(C_k|X) = \prod_{i=1}^{n} P(X_t|C_k) \quad (3.3)$$

Here, X_i are values for features from the available samples $'X'$.

Algorithm 3.1: Naive Bayes classifier

Input: Known and unknown samples to construct the learner model

Output: Prediction accuracy of testing data

Repeat

Dataset = randomize (known patients' sample);

T= 'N' bins generation from D;

For $j = 1\ toN$ do

Test = $T[j]$;

Train = $D - test$

Learning = learning (train, scheme)

Outcomes = Test_classifier (learner, test);

//Evaluate performance metrics of learner on testing

End

Till N times

Average results = $\frac{1}{M*N}\sum result$

Best strategy = choose (average results, learner);

//choose learner after processing

Finest Learning process = learning (Dataset, best results);

Final_prediciton accuracy = Test_classifier (finest learner, unknown sample classes)

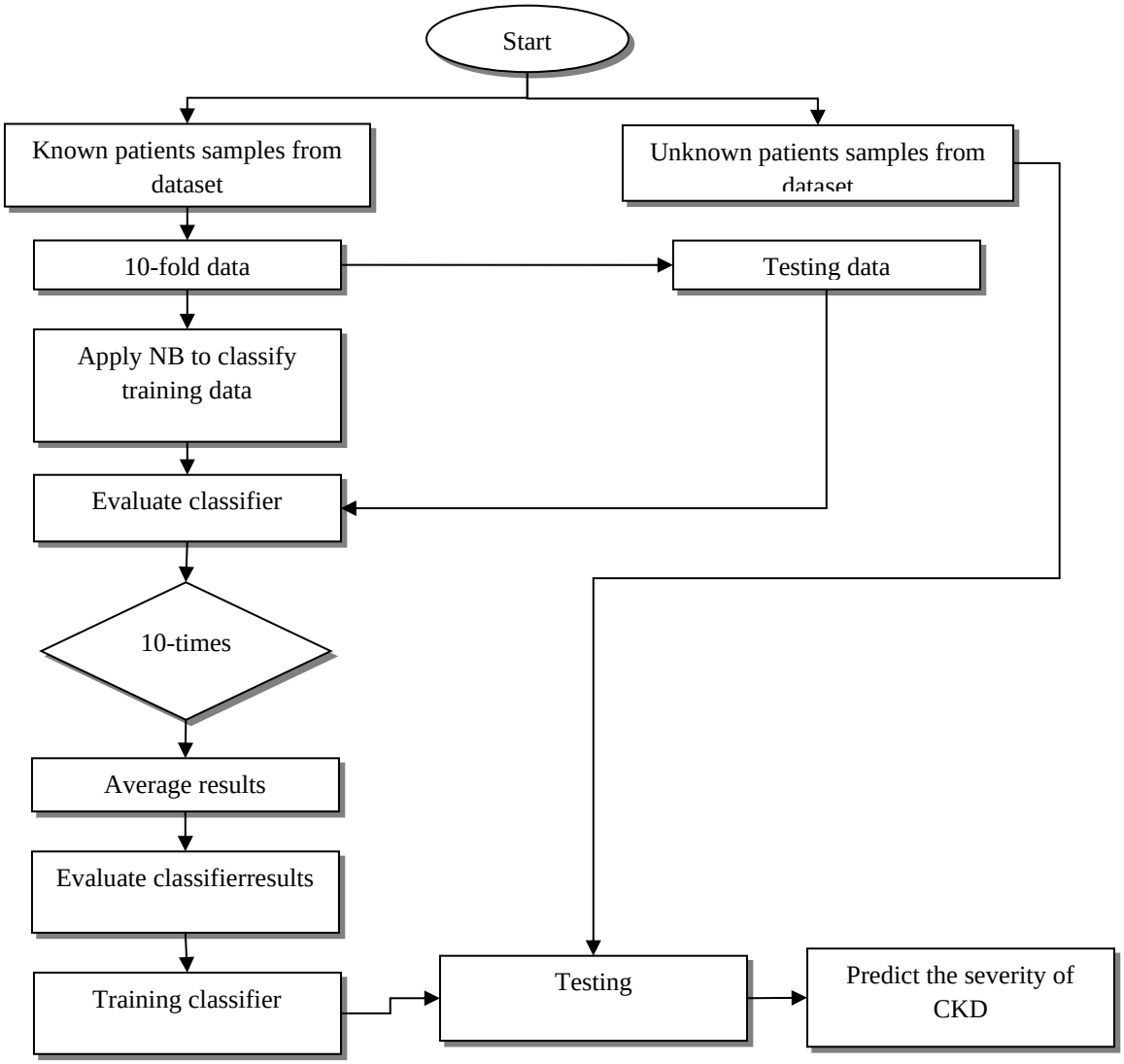

Fig 3.4 Flow diagram of NB predictor model

3.3.3. Choice Based Hierarchies (CbH)

CbH is an assumption that is used for construction of regression or classification problem in a hierarchical mode. It is partitioned into number of subsets and parallel generates leaf and choice nodes. Two kinds of CbH are utilized for extracting information is analyzed with

categorical and regression hierarchies. It is to evaluate the outcome and to fit the precise information class/real numbers. The classifier model is used to model a CbH as explained in Fig 3.6:

The hierarchical model is a resourceful environment to demonstrate the complex interactions among the stochastic model parameters. Specifically, it is generally utilized to demonstrate the population models, where the parameters are classified as an individual to establish the relationship among the parameters of every individual that comes under similar population. This chapter pretends to cope with the classification problems of CKD prediction of various patients with the measurements of features that are available. Probability measurements of these measurements are based on the patient's condition; moreover, all the patients who suffer from same disease are considered to be associated with one another based on probability distributions.

NB_CbH provides an effectual way to specify the association among the certain appropriate relationship by determining the conditional independence structure and are more appropriate to set the conditional probability distributions as in Fig 3.5. The structural model of this CbH is given as below: Assume, certain variables $'x'$ are measured as n_{rep} times for $'n'$ patients that come under same population, that is, they posses same CKD symptoms. Consider $'x'$ is stochastic variable that is based on parameter set θ, whereas for $'i^{th}'$ individuals, these dependencies are based on the probability measure of $p(\theta_i|\varphi)$, where, φ is hyper-parameter set of given population. Similarly, the subjects are classified by probability distribution based on population parameters for subjects of same population. Consider, prior distribution to φ, where $p(\varphi)$ is joint distribution with $p(\varphi, \theta) = p(\theta|\varphi)p(\varphi)$, where $\theta = \{\theta_1, \theta_2, .., \theta_m\}$.

When the dataset is $X = \{X_1, \dots, X_n\}$ is attained for all $'n'$ patients, where $X_i = (x_{i1}, x_{i2}, \dots, x_{inrep})$ is i^{th} patients' measurement, where joint posterior distribution $p(\varphi, \theta|X)$ is evaluated using NB. It is simpler to provide these distribution which is directly proportional to $p(X|\theta)p(\theta|\varphi)p(\varphi)$. The population factors are generally unknown, where the integrals over φ facilitates the evaluation of θ model parameters from the data of patients'.

Generally, this choice based hierarchy is applied over various contexts that range from signal processing to medical field, and pharamacology to bio-information. Similarly, various

computational approaches are specified for fitting this choice based hierarchy with discrete responses, multivariate, survival, and time series models. This hierarchical model is used over various NB model to relax the consideration over the conditional independence among the attributes.

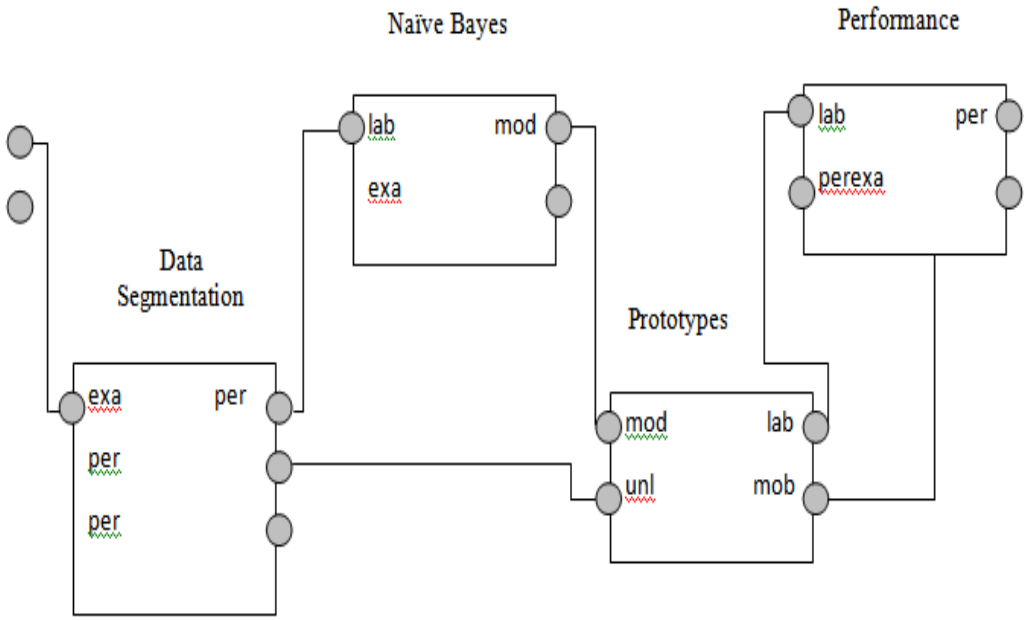

Fig 3.5 NB performance screen

3.4. Computation with CbH

From probabilistic insights, classification problem is considered as the class selection with higher posterior probability with available data. This work explicitly deals with the multiple replicate values. In case of CKD prediction, an individual is consider having similar symptoms and replicates the multiple tests for the patients. Hence, the computation of target (features) is provided with various evaluation of feature expression for all patients.

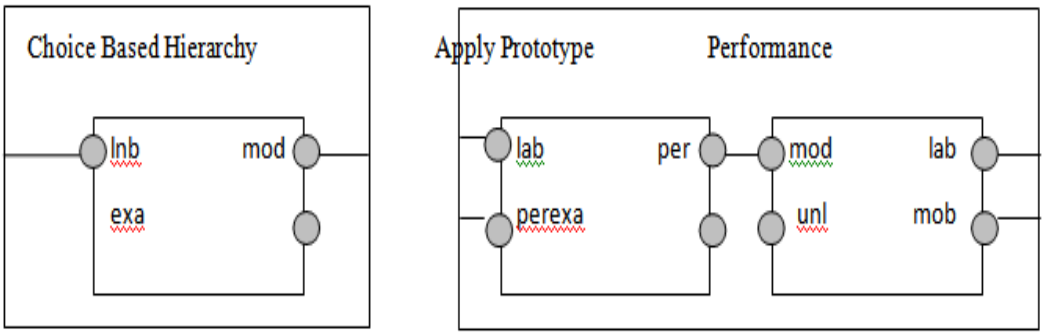

Fig 3.6 CbH model

Assume, x_{ij}^{lk} is duplicate measure of features of i^{th} individuals with respect to class $'k'$. For simplicity purposes, the given class $'k'$ and features are expressed as $x_{ij}, j = 1, \dots, n_{rep}, i = 1, \dots, N_{ck}$; where N_{ck} is number of individuals C_k. Consider, that the replicate values of generic cases are normal distribution with mean value μ_i with variance σ^2,i.e., independent of $'i'$,however dependent on $'k'$ class $x_{ij} \sim\sim N(\mu_i, \sigma^2)$. The mean value is normal distribution with the population, the mean with variance $\tau^2, i.e., \mu_i \sim \ll N(M, \tau^2)$. Variance is similar for all CKD patients who belong to same class by reflecting notion variability based on disease heterogeneity which is the property in disease prediction. This assumption is considered to be same for CKD prediction which turns out be evident, when evaluating the data variance. The reliability is increased by huge amount of exploited measurements. The structure of CbH is given in Fig 3.7.

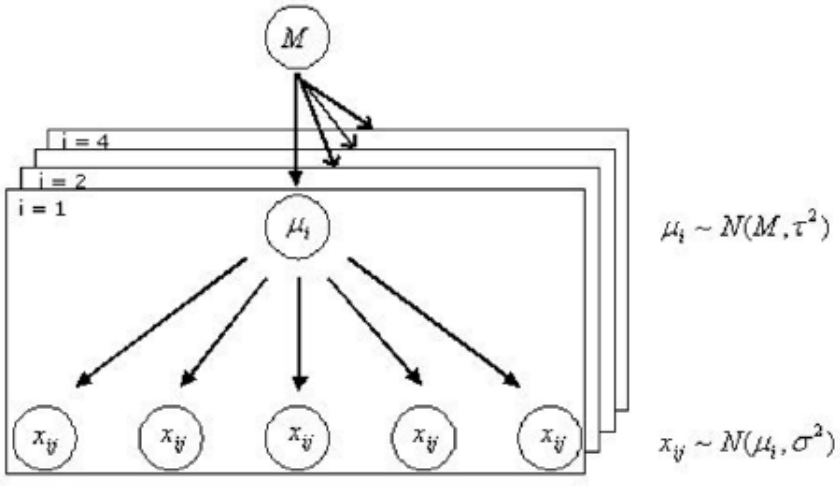

Fig 3.7 Structure of CbH model

With given probabilistic model, this work concentrates on classifying the newer classes by assisting the class model parameters like M, τ^2, σ^2 for all known classes. This model describes how to learn model parameters from dataset. However, learning phase and classification is reported based on the traditional NB method to show the differences. To categorize the newer cases in NB, it is essential to compute posterior probability of every class in provided data. The vector value is depicted as $X_i = (x_{i1}, x_{i2}, \dots, x_{inrepi})$ which specifies replicate measurements of cases from provided features (univariate cases). For simplicity purpose, the sub-index $'i'$ is avoided.

With the utilization of Bayes' theorem, posterior probability of class C_k is provided with a set of data as in Eq. (3.4):

$$P(c_k|X, \sigma^2, M, \tau^2) = \frac{P(X|C_k, \sigma^2, M, \tau^2)P(C_k)}{p(X)} \; \alpha \; P(X|C_k, \sigma^2, M, \tau^2)P(C_k) \tag{3.4}$$

To compute posterior probability, marginal likelihood $P(X|C_k, \sigma^2, M, \tau^2)$ is evaluated with conditional independence exploitation based on the hierarchical model which is expressed as in Eq. (3.5):

$$P(c_k|X, \sigma^2, M, \tau^2) = \int_{\mu} P(X|C_k, \sigma^2, M, \tau^2)P(C_k)\, d\mu \tag{3.5}$$

The marginal likelihood is provided for the sake of readability and the model parameters τ^2, σ^2 is avoided from the equation and expressed as in Eq. (3.6) & Eq. (3.7):

$$P(X|C) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{n_{rep}}(\tau\sqrt{2\pi})} * \int_{\mu} \exp\left(-\frac{1}{2\sigma^2}\sum_{j}\left(x_j - \mu\right)^2 - \frac{1}{2\tau^2}(\mu - M)^2\right) d\mu \tag{3.6}$$

$$P(X|C) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{n_{rep}}\sqrt{n_{rep}\tau^2+\sigma^2}} \exp - \left(\frac{\sum_j x_j^2}{2\sigma^2} + \frac{M^2}{2\tau^2}\right) * \exp\left(\frac{\left(\frac{\tau^2 n_{rep}^2 x_{mean}^2}{\sigma^2} + \frac{\sigma^2 M^2}{\tau^2} + 2n_{rep}x_{man}M\right)}{2(n_{rep}\tau^2+\sigma^2}\right) \quad (3.7)$$

At last, with the provided model parameters, newer class X is classified as class reduces posterior probability is directly proportional to marginal likelihood when class are priori equal to the model.

The significant functionality is the production of classification rule via marginal likelihood which comprises information with heterogeneity. These information are expressed as σ^2 which is provided with the decisions where there are some clear difference among the classes. The traditional approaches like NB and quadratic discriminant analysis is considered as the sample variability, which is expressed as in τ^2. However, the classification rules are evaluated in closer form, where it is utilized in real-time applications like CbH-NB classifiers.

The generalization of multivariate classes, that is, $\bar{X} = X^1, X^2, \ldots, X^{Nfeature}$ which is attained from the assumption of CbH-NB classifier, where the conditional independence of features in provided class is expressed as in Eq. (3.8):

$$P(\bar{X}|C) = \prod_{l=1}^{N_{feature}} P(X^l|C) \quad (3.8)$$

Here, the posterior probability of $'k'$ class is expressed as in Eq. (3.9):

$$P(C_k|\bar{X}) = \frac{P(\overline{X}|C_k)P(C_k)}{P(\overline{X})} \; \alpha \prod_{l=1}^{N_{feature}} P(X^l|C_k)P(C_k) \quad (3.9)$$

The attained results are based on the computation with the provided data and the classification is done with appropriate classification of classes with CKD and non-CKD respectively. Fig 3.8 depicts the prediction of newer classes using CbH-NB classifier.

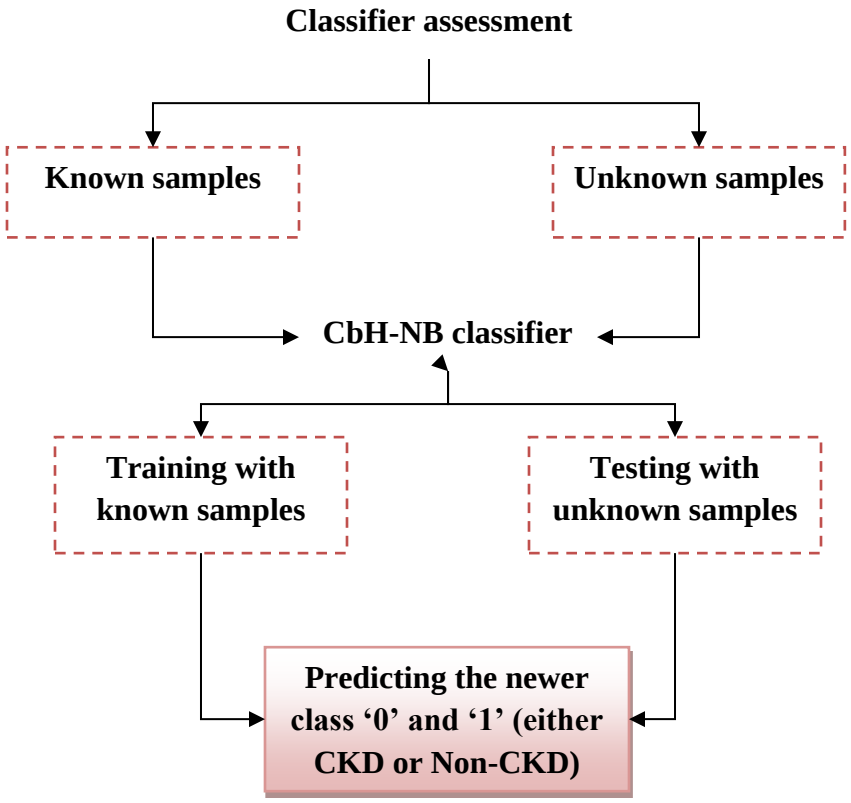

Fig 3.8 Prediction of newer classes using CbH-NB classifier

3.5. Summary

The overall target of this work is to analyze the dominant features associated with the diabetic dataset and to extract the information with diverse kinds of medical information. The CbH functionality is provided with 90.2% precision when computing with NB classifier. The classification scheme for the diabetic information set is attained with proper class extraction. It is associated with the extraction to revive synchronization from elements are not represented with categories for evaluation purpose. It is concentrated on evaluation with enhanced system performance assessment where NN precision and clustering is based on class evaluation.

CHAPTER 4

DETECTING CHRONIC KIDNEY DISEASE THREAT LEVEL

4.1. Prologue

Generally, CKD is provided by a fact that, 'kidney losses its regularized features over certain time due to various external factors. The earlier prediction and treatment of CKD helps to save kidney and eliminates the CKD progression. It is a global public health issue over past few decades. Risk associated with CKD in the developing countries is undergoing the therapy is costly. The CKD prediction strength relies on threat prediction which causes grouping schemes which cannot be under-rated as it varies the series of ailments. Initially, the phase of predicting the fatal CKD is originated from the chances for performing probable policies in minimizing the probability of kidney related issues. Here, Neural Fuzzy scheme is used for predicting the threat. This model overcomes issues associated with NB classifier with CbH. The constraints associated with NB are the assumption of independent predictor features where the attributes are mutually independent. Here, the predictor set completely independent to one another. In order to overcome these issues during the prediction of threat, this work concentrates on the enhanced Neural Fuzzy model and Random Forest approach to enhance the constructing of disease phases.

4.2. Research Objectives

The target of this work is:

- ✓ To design a Neural Fuzzy model for predicting the threats over the model and to perform detection of CKD.
- ✓ To apply the Random Forest characteristics to reduce the complexity during the execution of the prototype model.
- ✓ To predict the threat parameters associated with the given diabetic dataset
- ✓ To analyze the functionality of Adaptive Neuro-Fuzzy Interference scheme to eliminate the drawbacks identified in NB classifier.
- ✓ To extract confusion matrix and other performance measure

4.3. Research Methodology

Various Machine Learning approaches are used for constructing the effectual prototype for predicting Chronic Kidney Disease (CKD) where certain performance metrics are used for evaluating the metrics like specificity, precision, and sensitivity of prototype. Before applying certain ML approaches, there are some essential pre-requisites have to be carried out for categorizing the pre-dominant elements. A characteristic classification scheme known as Random Forest (RF) is used to perform the selection of pre-dominant features. The ultimate intention is concentrated on utilization of ML approaches known as ANFIS systems termed as unsupervised learning. These approaches are used for predicting the classes of CKD patients which fit with certain specifications. Fig 4.1 depicts the block diagram of proposed ANFIS model.

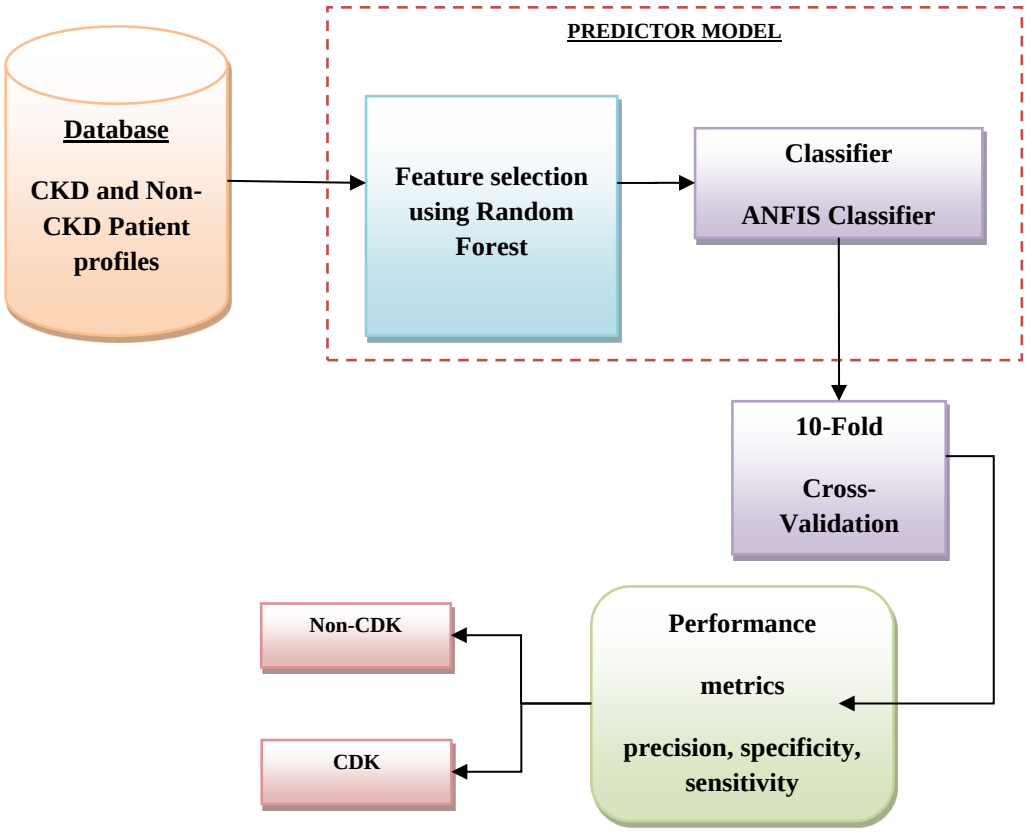

Fig 4.1 Block diagram of proposed ANFIS model

4.4. Formulating Problem

From the literature, it is an extensive concern where higher attention is given to the requirements towards CKD connectivity and other related diseases with grouping and risk evaluation. The results attained after completing the evaluation provides the association among the health policies for handling CKD prediction effectually. Fig 4.2 depicts the relationship of CKD with other diseases such as diabetics. Individuals with these kinds of CKD have enhanced the threat prediction which is related to the diseases.

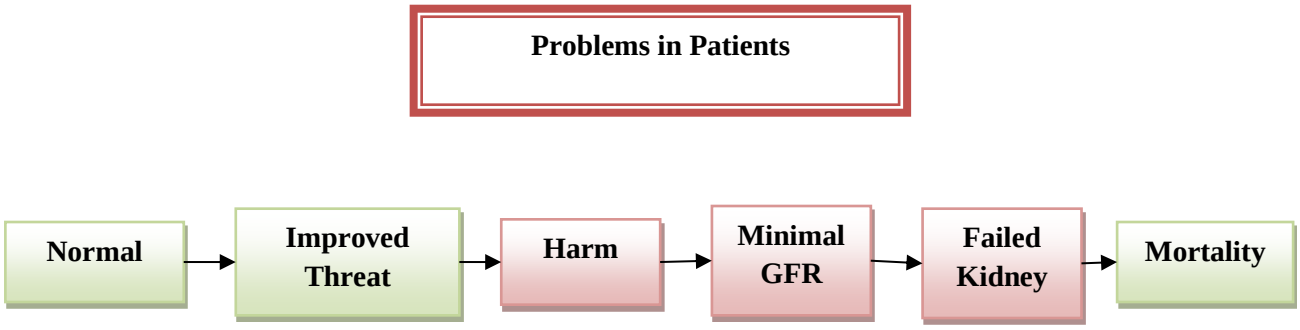

Fig 4.2 CKD possess threat characteristics to attain certain undesirable results for various other diseases

4.5. Techniques

The target of this work is to model a prototype to categorize patients during emergency condition by generating smart predictor for the nephrologists. The predictor model is a phase where the system is applied for predicting the emergency condition of the patients by measuring certain clinical parameters. This predictor model is the combination of NN and ANFIS is classified by controlling complicities and characteristics measure. This model pretends to reduce the execution time for examining the emergency patients' condition waiting in the queue. The design intention is to use an effectual ANFIS is not used to measure threat during the classification process. The medical information is extracted from the available IFTS data sheet from emergency department that comprises of text, classification and some constant information. The outcomes of precision are based on ANN that provides ANN functionality in an effectual manner as that of ANFIS using

the triage analysis. It is known that, the anticipated ANFIS model is more effectual for the evaluation of available data and the other unknown information.

The modeling of a healthcare prototype relies over the available IFTS data sheet which includes the voice pathology analysis. Moreover, there are two diverse characteristics that include speech signal and MPEG – 7 minimal audio levels with interlaced derivation pattern which is utilized for processing speech and voice signals. Here, the characteristics attained from Gaussian prototype are also used as categorizers. Also, it comprises of VPA system for examining the efficiency based on the pre-requisite execution time and precision.

Based on the advancements encountered in IoT, the genomic information, e-health records, public-health based records and behavioral information for generating enormous information is considered to be healthy with hidden patterns of big data analytics are capitalized to extract the perceptions and proof for reducing average healthcare cost as it produce effectual outcomes based on smart device usage.

Based on the analysis, it is observed that there are various types of CKD which is classified based on the type of severity. The simulation is done in MATLAB environment for analyzing the characteristics of ANN and SVM with the available tool box. The ultimate target is to compute the performance of the model for predicting CKD based on the accuracy and precision values.

The metrics are generated and analyzed for measuring the severity of CKD using MATLAB tools. Analysis is concentrated on the usage of various schemes relies on the available MATLAB and evaluating the outcomes of other tools. Here, ten-fold CV is used for classifying the patient's class using the analyzed characteristics. This method is considered to be more obvious based on RF for pre-processing and results are evaluated based on these techniques.

The goal is to analyze the target of CKD prediction model for diabetic patients with the use of various ML approaches. Here, certain recommendations are attained from CbH to retrieve effectual results with the desirable precision values by computing the performance based on sensitiveness and specificity. Here, MATLAB tool is used for analyzing the diabetic dataset using RF and ANFIS approaches where RF is applied for pre-processing the data and ANFIS is applied for classifying the patients' class.

The execution for analyzing the threat with the developed CKD predictor model along with the use of healthcare records evaluation is based on time-series analysis. Here, the goal is to resolve the drawbacks encountered in heterogeneous dataset for evaluating the threats over the stages of CKD, i.e., 3 to 5.

The analysis over the features of CKD is performed; based on this analysis various tools are employed for simulation purpose. The target is to extract the practicable solution using MATLAB and WEKA tool. It verify and validate the CKD evaluation as it is used for performing comparison against R programming for superior information and simple big-data combination like Hadoop and Mongo respectively. Here, reviews are done with schemes like logical regression, SVM and multi-layer perceptron with prevailing approaches.

The objective is to improve CKD prediction with MATLAB for executing ANFIS that works effectually than existing probability-based approaches. This work helps in addressing the issues connected with patient's life and uses them to reduce precision against the anticipated statistical scheme. Various existing approaches make use of computerized disease prediction model using ANFIS and Artificial Neural Networks. It is performed with MATLAB environment by providing the inputs from UCI Machine Learning Repositories. The predictor model has to examine the severity of renal failure and to predict it in its preliminary stage. In this research various classification approaches like ANN, NB, CbH, and ANFIS are used for predicting the renal failure.

The above-mentioned approaches are used for predicting the renal failure where the objective is to predict CKD earlier. The most dominant features have to be selected to perform further classification and to predict the stages of CKD of individual patients'. The anticipated classification prototypes with various approaches comprises of concealed prototypes and optimal feature selection for evaluating and classifying CKD patients. With the attained outcomes, it can provide enhanced precision over the dataset utilized for classification of data from the dataset as the inputs are attained from feature selection approaches. The automated disease predictor system is merged with Fuzzy Logic and NN known as ANFIS which is used for evaluating the threats over the CKD prediction. For automating CKD evaluation, various iterations are performed. The results demonstrate that the recommender model is extremely suited for predicting the higher and lower chances of cardiac risks for the CKD patients.

With the evaluation and examination, the purpose of this work relies over the benefits of the healthcare domain where ML approaches are used for analyzing the possibility of the disease. The research gaps need to be bridged and to be resolved by other techniques which intend to design ANFIS model for predicting CKD with the grouping of results to identify significant classes of diabetic individuals related to the CKD outcomes. The characteristics and the features related to the high risk CKD patients' are analyzed with other chronic disease like diabetes is extremely complex. It is due to the fact that CKD is highly connected with various other complications and some other diseases.

4.5.1. Adaptive Neuro-Fuzzy Inference System (ANFIS)

The clustered object that comes under an identical class is known as cluster. These objects are grouped into one group and other objects are considered under other objects. The clustering process is to produce identical class objects by coordinating the objects over the cluster. The cluster evaluation comprises of partitioning the dataset into clusters with resemblances against label allocation to generate various clusters. The clustering process is more flexible to change and assists in developing helpful features that differentiates the cluster.

The integration of Fuzzy Logic and Neural Network is described as an efficient system that is skilled with the use of precise learning process that is attained from ANN foundations. The Learning is done with the data while the variations are performed with a local means. This work integrates ANN with FIS. The inference system is constructed over the ANN framework and the learning processes are used to accelerate the FIS metrics that are utilized as bias function. ANFIS strategy is based on implementing neural fuzzy on the UCI machine learning repository.

4.5.2. Neuro-Fuzzy

This approach is well-known scheme where the Artificial Intelligence (AI) domain is utilized to predict appropriate information in every domain that is related to it. NFS function is based on training is tuned based on bias function by merging Fuzzy system. NFS is utilized to adjust bias function. It is alike of ANFIS and works like Fuzzy inference system based on its functionality. It is based on the Tsukamoto and Sugeno prototypes. This hybrid scheme is adjusted using ANFIS. It is specified with the fact that it comprises of two inputs $'x'$ and $'y'$ and the

outcomes$'z'$. Based on this prototype model, the Sugeno Fuzzy model is produced with the following rules. Fig 4.3 depicts the block diagram of Fuzzy system model.

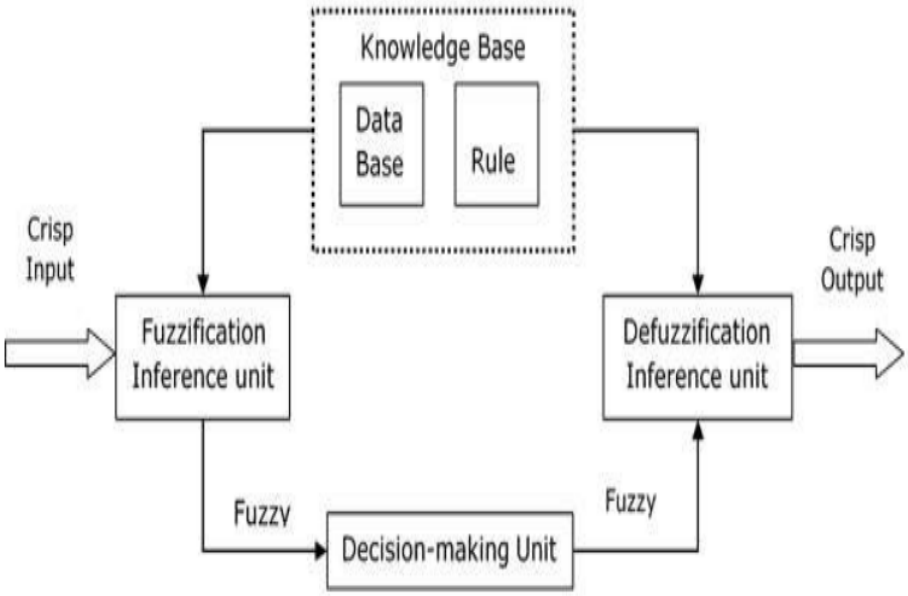

Fig 4.3 Fuzzy system

This model is designed by Takagi, Sugeno and Kang (1985). The rule format is provided as:

$$IF\ x\ is\ A\ and\ y\ is\ B\ THEN\ Z = f(x, y)$$

Here, AB are fuzzy sets in antecedents while $z = f(x, y)$ is a crisp function. FIS process under TS fuzzy method functions in the below given form:

Step 1- Inputs fuzzification – Here, system inputs are made fuzzy

Step 2 – Using Fuzzy operator – Here, fuzzy operators are used to attain the output.

Rule format

Sugeno rule format is provided by an example below:

$$if\ 7 = x\ and\ 9 = y\ then\ output\ is\ z = ax + by + c$$

Rule 1

If $'x'$ specifies r_1 and $'y'$ specifies s_{p1} then $c_1 = a_1x + b_1y + k_1$

Rule 2

If $'x'$ specifies r_2 and $'y'$ specifies i_2 then $c_1 = a_2x + b_2y + k_2$

The working functionality of ANFIS framework is explained as:

ANFIS network model comprises of two nodes: adaptable and fixed. These layers are composed of various other nodes predicted from node function. The anticipated model efficiency is based on adaptable parameters over nodes. The network learning functionality is based on certain parameter settings for reducing error over the appropriate output. It is depicted in Fig. 4.4 with two inputs and one output. The afore-mentioned rules are based on FIS method (Takagi-Sugeno) form.

i) First Layer

This layer s termed as fuzzification layer where every node is considered as an adaptive membership function. The parameterized functions are membership function for fuzzy set or linguistic label which is of trapezoidal generalized bell or gaussian/triangular form, i.e., Gaussian membership function is depicted based on parameters couples (c, σ).

$$y_i^{(1)} = Gaussian\,(x: c, \sigma) = e^{\left(-\frac{1}{2}\right)\left(\frac{x-e}{\sigma}\right)^2} \tag{4.1}$$

Here, gaussian membership parameters are regularized by c, σ, these parameters are pointed as antecedent parameters, $y_i^{(1)}$ is output layer.

ii) Second layer

This layer is known as product layer, where every nodes can reacts tosugeno fuzzy rules. Nodes are gathered from input with respective to neuron fuzzification. As an outcome, the firing strength of every rule is presented. The neuron output obtained from this latey is specified as in Eq. (4.2):

$$y_i^{(2)} = \prod_{i=1}^{k} x_{ji}^{(2)} \tag{4.2}$$

Here, $x_{ji}^{(2)}$ is input layer from $1(j)$ to $2(i)$ layer and output is specified as $y_i^{(2)}$ for neuron $'i'$ in product layer.

iii) Third Layer

It is a standardized layer. It accepts feedback from product layer (neurons) and evaluates the weighted firing power. The neuron results are expressed as in Eq. (4.3):

$$y_i^{(3)} = \frac{x_{ji}^{(3)}}{\sum_{j=1}^{n} x_{ij}^{(3)}} = \bar{\mu}_i \tag{4.3}$$

Here, $x_{ij}^{(3)}$ is received input and neurons generated from product layer to neuron over normalization layer $y_i^{(3)}$ is layer 3 output.

iv) Fourth Layer

It is a defuzzification layer. The nodes are considered as adaptable or modifiable nodes. Neuron over defuzzification layer computes corresponding weights and values of certain rules are specified as in Eq. (4.4):

$$y_i^{(4)} = x_i^{(4)}[k_{i0} + k_{i1} + k_{i2}] = \bar{\mu}_i\,[k_{i0} + k_{i1} + k_{i2}] \tag{4.4}$$

Here, $x_j^{(4)}$ is input of layer 4 while output is $y_i^{(4)}$. $k_{i0}, k_{i1}, k_{i2}$are successive parameters of $'i'$ rule.

v) Fifth Layer

This layer gives complete output for ANFIS model that integrates the output of previous layers. It is expressed as in Eq. (4.5):

$$y = \sum_{i=1}^{n} x_i^{(5)} = \sum_{i=1}^{n} \bar{\mu}_i [k_{io} + k_{i1} + k_{i2}] \tag{4.5}$$

ANFIS learning process comprises of various updated parameters using two-pass learning, backward/forward pass algorithm. ANFIS parameters are trained for reducing the error among the desired and actual output as depicted in figure given below.

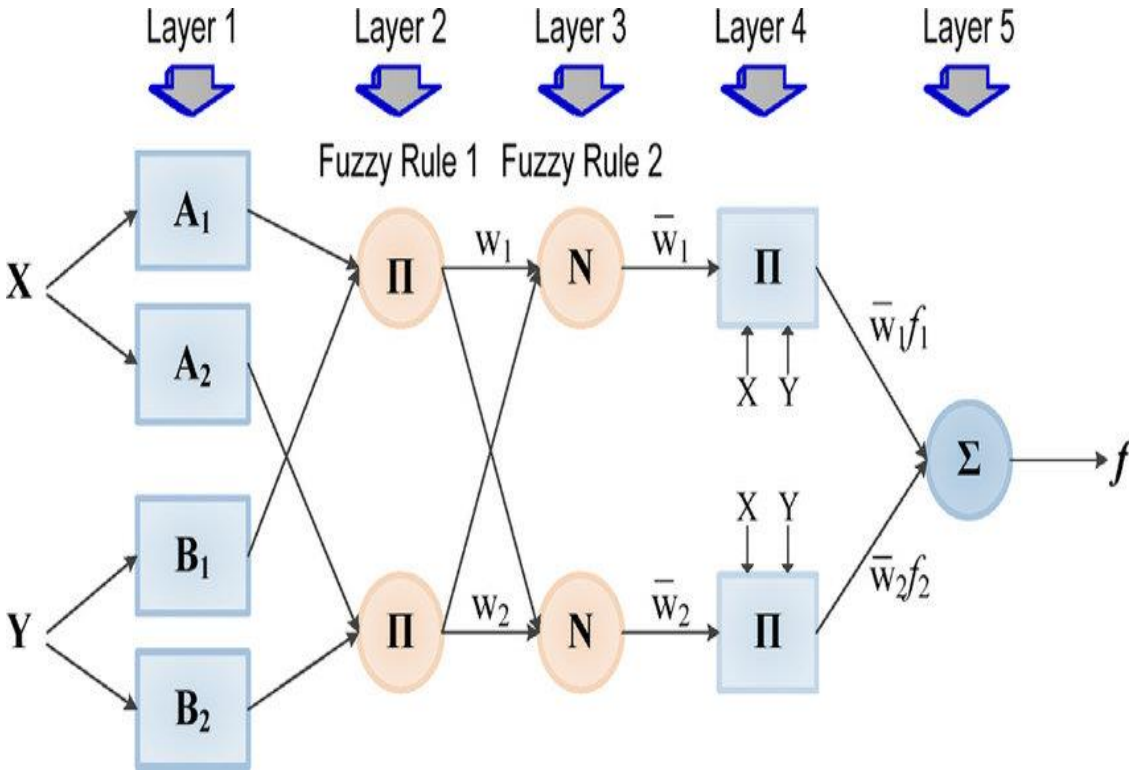

Fig. 4.4 Neuro – Fuzzy Framework

4.5.3. Tree-Based clustering

For all cases of CKD dataset, the samples are considered as a group/cluster. These are aggregated until the entire group is generated. Here, $b = \{b_1, b_2, b_3, \dots, b_n\}$ are retrieved from CKD datasets.

1) initially, the disjoint group are initiated from level '0' and the successive sequences are '0'.

2) The next step is based on the predicting minimal distance that exists between pairs from the available groups. For example, the pairs $(a), (b)$ are based on the distance among $d[(a), (b)] = \min d[(x), (y)]$ and the minimal distance of the existing groups are also considered.

3) The sequence number is increased as $s = s + 1$. Clusters (a) and (b) are merged as single group generated from the successive group $'g'$. Grouping levels are sets as $1(s) = d[(a), (b)]$.

4) From the CKD dataset, distance matrix is revised and specified with $'d'$. It is attained by eliminating rows and columns comes under the clusters (a) and (b). However, the rows are aggregated into column that is associated with the newer cluster. The distances among the newly generated clusters are specified as (a, b) and the existing clusters (g) are specified as: $d[(g), (a, b)] = \min(d[g), (a)],\ d[(g), (b)]$.

5) Finally, steps 2 to 4 are repeated until the groups are merged as an entire patients' group, else stop.

4.5.4. Random Forest

Originally, RF is modeled by Breiman which is extensively applied in both regression and classification analysis without tuning the hyper-parameters. This method is used for training number of DT predictors and averaged to reduce the over-fitting problem and to improve the accuracy. Moreover, it is well captured non-linear association patterns among the predicted and predictor variables. This model has the ability to attain un-biased deviations of classification and regression model for score estimation. Fig 4.5 depicts the RF based tree model of feature selection.

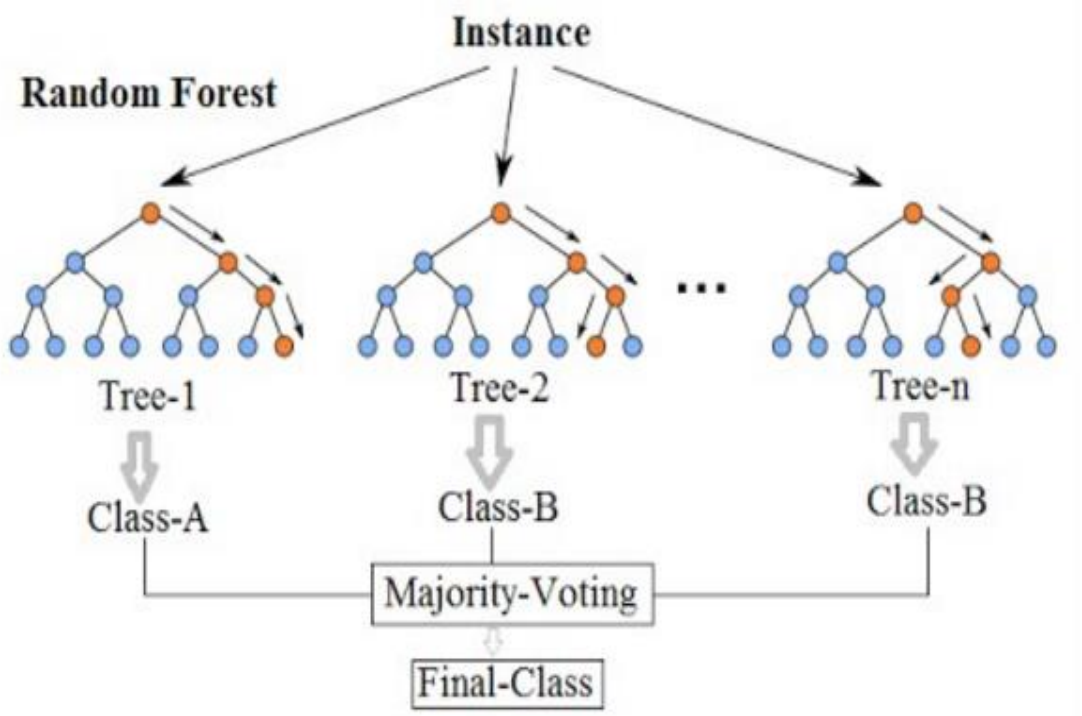

Fig 4.5 RF based tree model

RF is an ensemble model that utilizes bagging as ensemble approach and DT as individual model. This model is based on the ensemble model to attain better prediction accuracy and also to provide higher prediction score with the predictor values which is used for training the random forest variables. With the advantages mentioned, the mentioned model faces some drawbacks in interpretability and mathematical theory which makes it impossible to show the decision made by the model.

Algorithm 4.1: Random Forest
Choose ′n′ random subsets from the available training set Train the given ′n′ samples One random sub-set is utilized for training Optimal split of all tree based on random subset features, i.e., assume 10 features and select 5 features randomly from the 10 features

The individual tree predicts candidates/records in test set independently Make final prediction

From the above algorithm, it is known that ensemble classifier is extremely powerful which can be used for both regression and classification problem. Similarly, other individual models such as SVM can be applied with boosting or bagging ensemble to acquire better performance.

4.5.5. Design Policy

Fig 4.6 depicts the ANFIS system model that performs analysis, threat detection and clustering for CKD dataset.

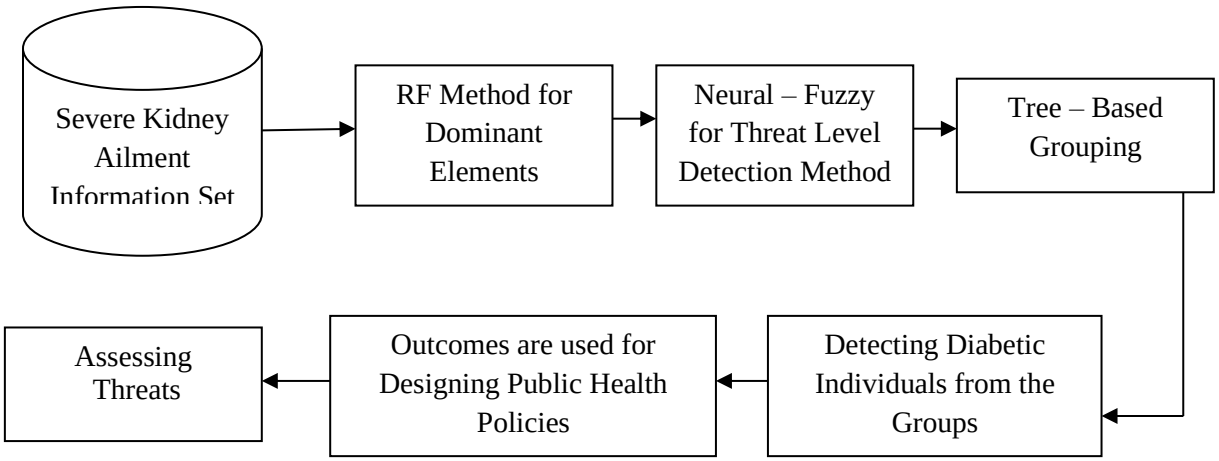

Fig 4.6 Design Policy

4.5.6. Design Planning

With the available big data analysis; it is probable to use improved patient information to include appropriate interference to patients at appropriate time. The process design is targeted to mine required elements with the use of RF characteristics. With the use of this appropriate characteristic information is considered as a threat level detection using ANFIS. The model used here to classify the patients-based on diverse CKD phases using symptoms and indications. The

results from ANFIS are clustered based on the tree-based clustering for predicting the essential clusters.

The data attained from UCI ML repository database comprises of 25 elements and 400 instances that makes use of above-mentioned schemes. In some cases, R tool is utilized for analyzing threat level for CKD disease prediction.

4.5.7. Data collection

With the above-mentioned problem statement, this work concentrates on analyzing the appropriate data from the UCI ML repository. CKD dataset comprises of huge amount of essential variables are more vital and essential for predicting the CKD severity in patients. Initially, dataset is in ARFF transforms suitable comma based value to be used in various programming language. Table 4.1 depicts the dataset variables and attributes related to CKD prediction.

Table 4.1 Dataset attributes

Elements	Description
Age (Years)	Years in Number
Blood Pressure	mm/Hg
Specificity Gravity	Range 1400 to 1000 (increases based on the risk)
Albumin	Range from 0 to 5 (higher is better)
Sugar Level	Level 5 represents the severity
RBC	Normal or Abnormal
Pus Cell	Normal or Not Normal (Increased number leads to urinary tract)
Pus Cell Clumps	Present or Not Present
Bacteria	Present or Not Present
Blood Glucose Level	mgs/dl
Blood Urea	mgs/dl
Serum Creatinine	High is not good
Sodium	mEq/L
Potassium	mEq/L

Haemoglobin	Less than 15 leads to kidney failure
Packed Cell Volume	Number
WBCs	Number
RBCs	Higher or less than normal
Hypertension	Yes or no
Diabetes Mellitus	Yes or no
Artery Disease	Yes or no
Appetite	Yes or no
Pedal Edema	Yes or no
Anemia	Yes or no
Class	CKD or not CKD

4.5.8. Pre – Processing and feature extraction

For attaining appropriate result, essential amount of pre-processing is needed. Most health-care information are generally lacks in missing values, noise and various other conflicting data. Some data are partitioned into equal amount of information, low-quality information generated from higher level ML results. Information refined and missing values are shifted with column values. Information extraction comprises of 25 instances where 14 nominal and 11 numerical values which are considered as initial CKD stages. Out of 400 dataset instances, 150 samples are non-CKD and 250 are CKD respectively.

RF based feature selection is evaluated for diminishing feature set of given prototype model. From Fig 4.6, it is known that RF is provided with the necessity for all characteristics with the use of Gini and Precision. Precision is defined based on the effects of every characteristic in the dataset with precision prototype and thereby unnecessary variables shows minimal effect over ANFIS prototype. With the mean of reduced Gini, the variables are represented by certain elements that remain over the top of some instances specified by points on x-axis and apply impurity in terms of Gini in each node based on the hierarchy expansion. Initially, ten features are used by ANFIS performs bias function in every feature as in Fig. 4.7. Fig 4.8 depicts the Gaussian bias function.

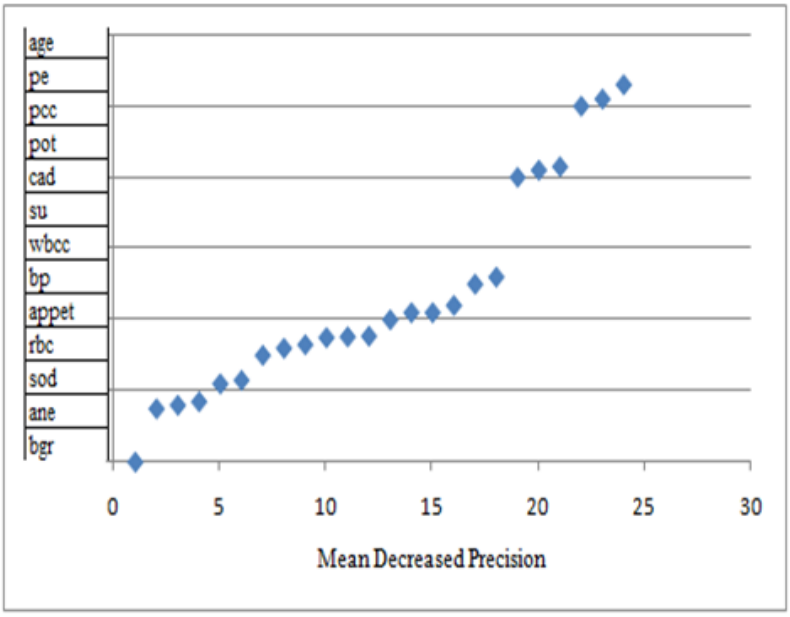

Fig 4.7 Mean precision value

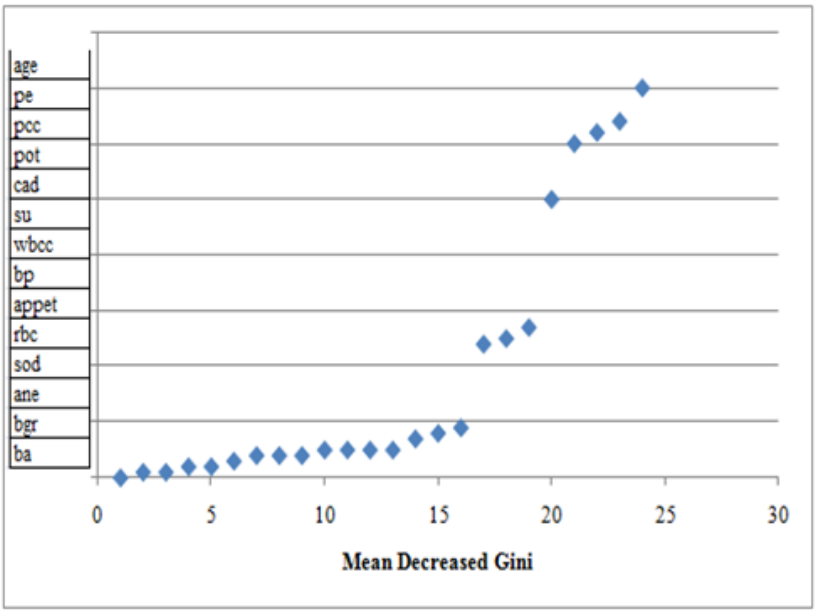

Fig 4.8 RF Characteristics Selection

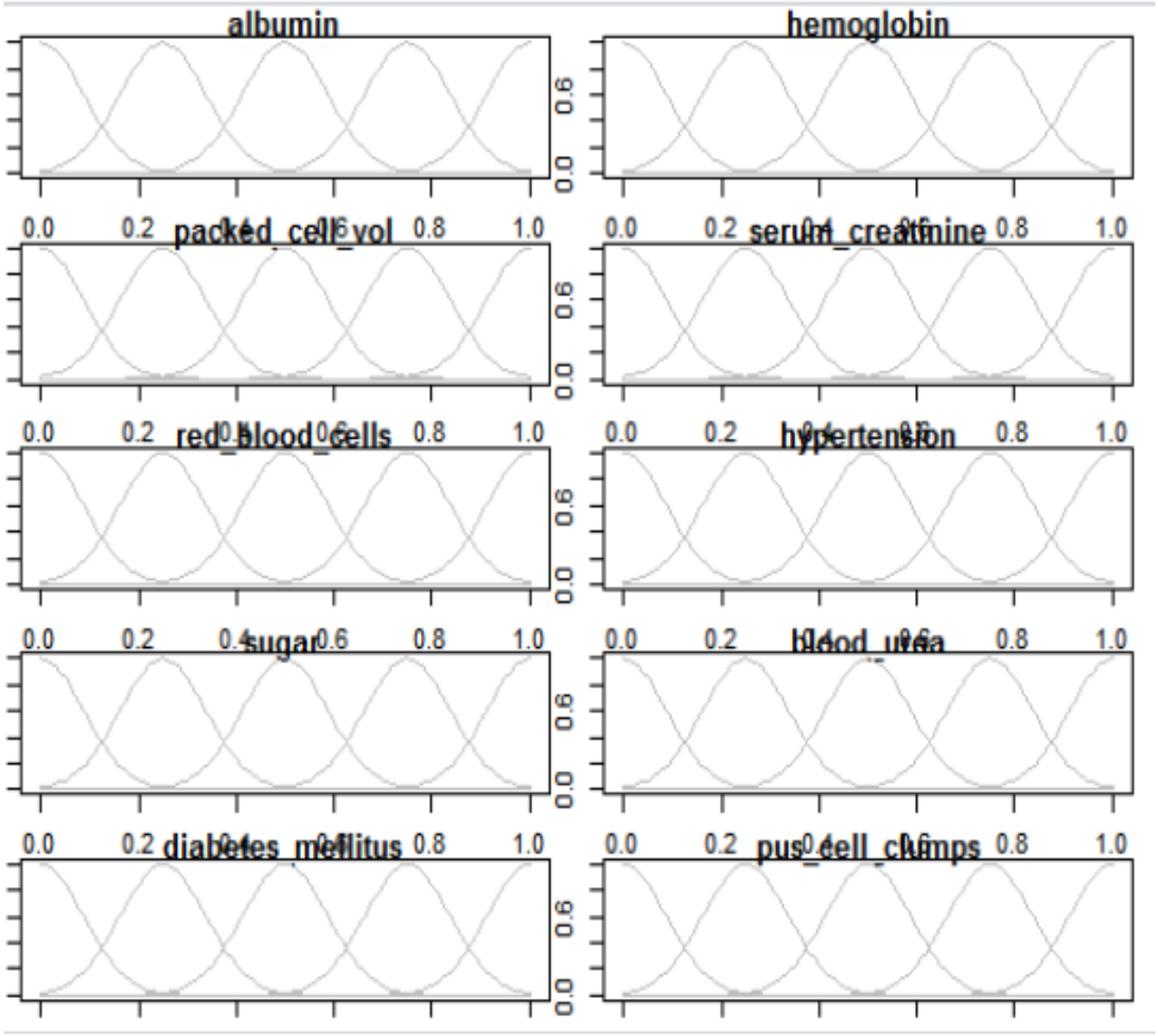

Fig 4.9 Gaussian Bias Function

4.5.9. Tree-Based clustering with ANFIS

With the application of elbow approach, it is known that some feasible amount of clusters from ANFIS predicts the appropriate results as in Fig 4.9. It is known that positioning of 'k' is based on clustering by setting the value as 3. It facilitates the prediction of diabetes with other related diseases.

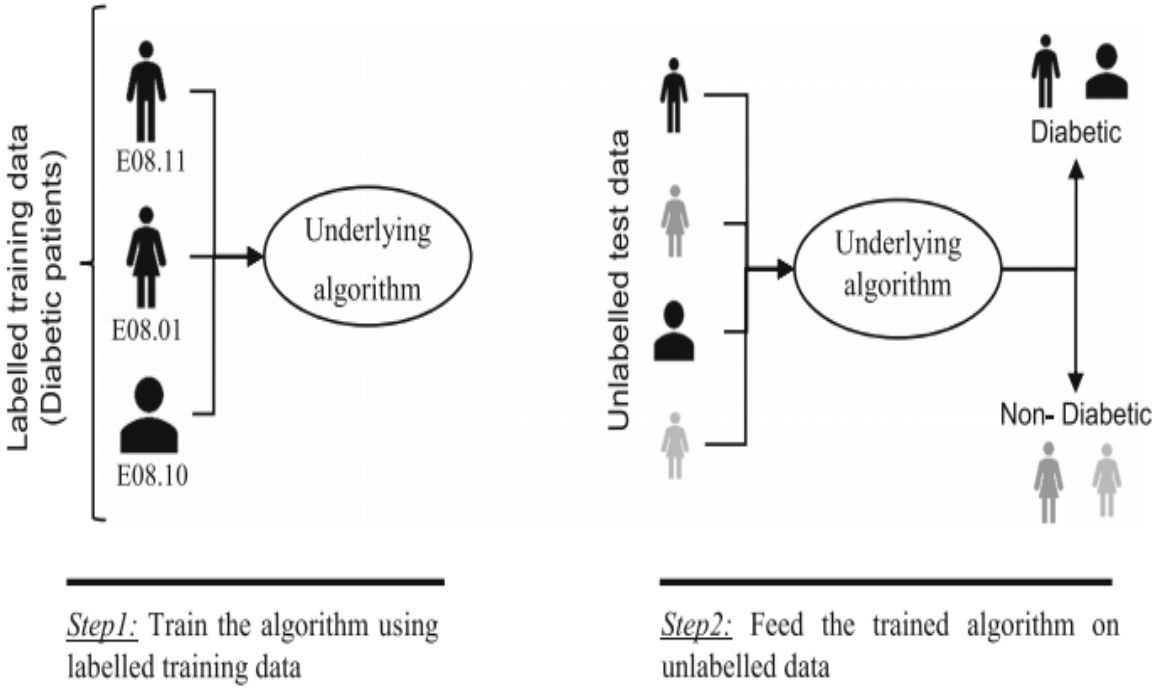

Fig 4.10 Classified Diabetic and Non-diabetic patients for CKD prediction

4.6. Summary

This section discusses in detail about the threat level over the designed prototype using the ANFIS model. Here, Random Forest is used for analyzing the features of diabetic patients to analyze the chances of CKD or not-CKD. The rules are generated and the features are grouped to provide an optimal outcome. The proposed ANFIS model pretends to give better solution than the existing model. Here, precision and Gini index are considered as the performance metrics. The precision value is considerably higher to forecast the functionality of ANFIS model. The feasibility of the proposed model is also verified by providing appropriate data for analysis. This model gives better trade off while compared to other models.

CHAPTER 5

NUMERICAL RESULTS AND DISCUSSIONS

5.1. Prologue

The ultimate goal of this work is to model an effectual CDSS for CKD using Machine Learning approaches. The primary target of this work is to consider a CKD benchmark dataset (UCI Machine Learning repository) that concentrates on designing predictor model for CKD prediction. The successive objective concentrates on modeling an efficient feature selection method to enhance classification accuracy with reduced feature set. The selection of feature selection and classification is based on an extensive study over the literature for modeling CKD predictor for medical application. The performance given by the proposed model is based on testing the UCI Machine Learning repository over the machine learning approach. The simulation is done on MATLAB R2013a that runs on 64 bit Windows operating system, Intel I3 processor, 2 GHz speed on 8 GB RAM.

This section discusses about the numerical results attained after testing the dataset over the proposed model. The classification results of this model is analyzed and compared with prevailing approaches. The performance metrics like accuracy, precision, F-measure are computed and evaluated.

5.2. Performance evaluation of NB-CbH model

This prototype model designed using machine learning approaches are validated to acquire higher precision. The subset is analyzed with the classifier model. Hence, the ratio of dataset partitioning is attained. With the use of NB classifier, the expected outcomes are attained for predicting CKD in earlier stage. The primary components are positioned for evaluating the features based on NB classifier model is age, gender, alcoholic habits, smoking, cholesterol HDL and so on. Fig 5.1 depicts the scattering points plotted for measuring the average age of individuals during the measure of CKD symptoms.

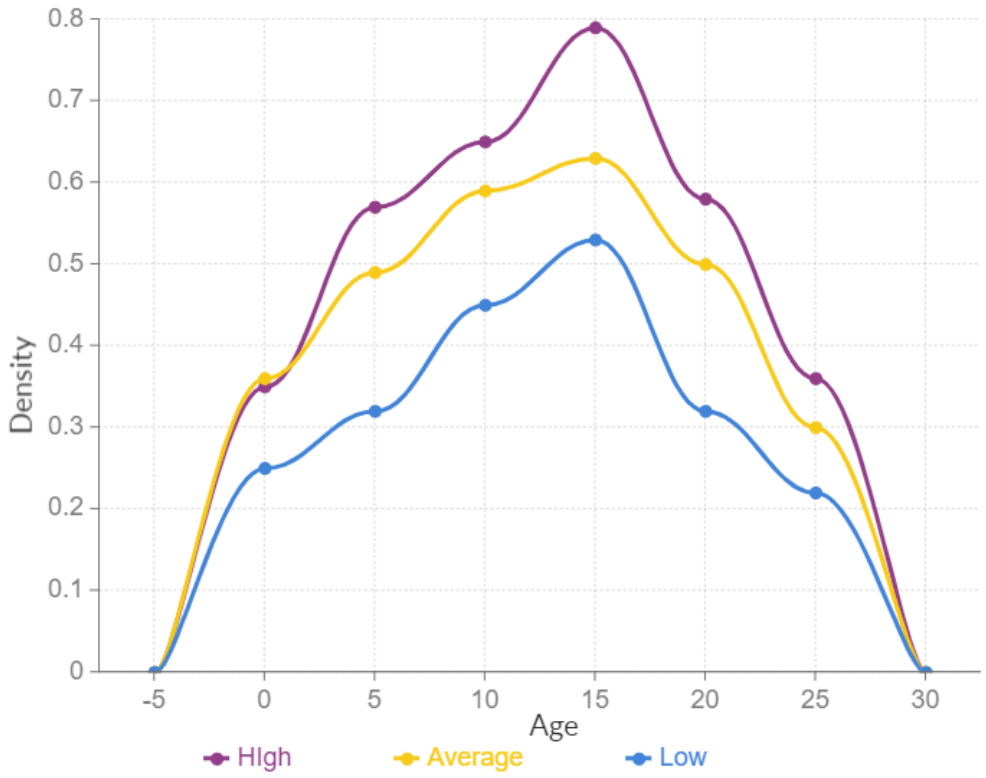

Fig 5.1 Bayes Scattering for Age

Here, the threat level is denoted as low, average and high values for certain age limits (50-55) which is specified in $'x'$ and $'y'$ axis respectively in Fig 5.1. Here, x-axis specifies age as a variable and y-axis specifies density. With estimated values the components related to age are considered and it is provided in the classification form. In Table 5.1, the mean and standard deviation of age feature is analyzed. The mean age group values are 52.25, 51.75 and 50.52 for low, age and high threat level respectively. Similarly, the standard deviations of these three levels are 10.05, 9.88, and 9.56 respectively.

Table 5.1 NB scattering representing Age

Elements	Metrics	Low	Average	High
Age	Mean	52.26	51.76	50.53
Age	SD	10.06	9.89	9.57

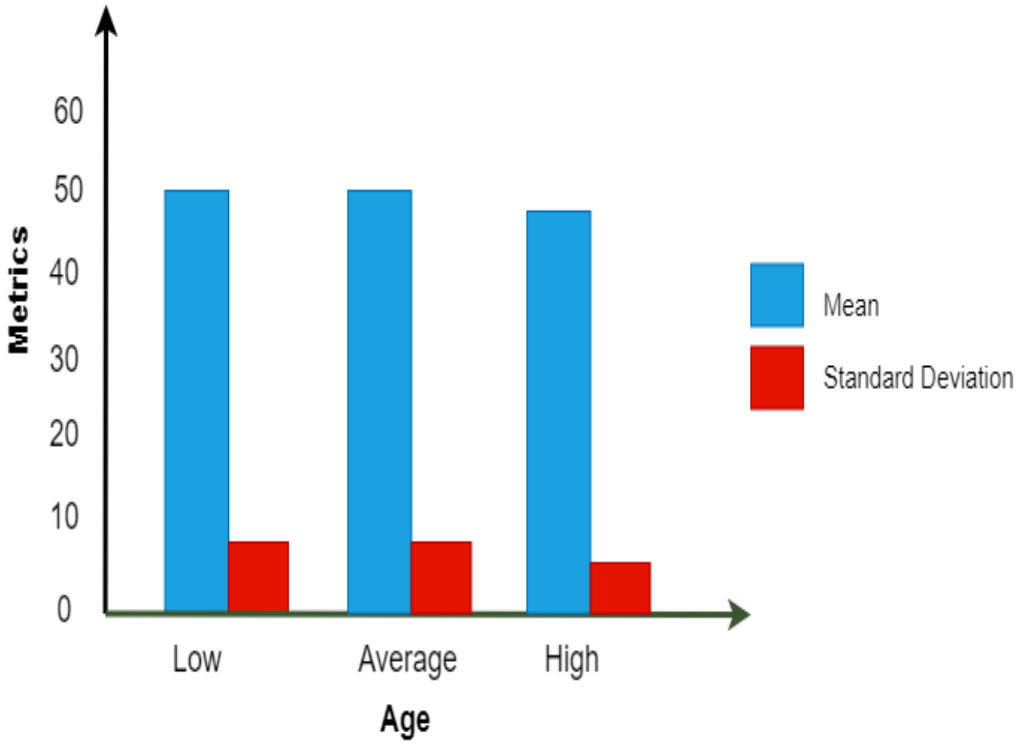

Fig 5.2 NB Scattering Table for Age

Fig 5.2 depicts the threat level for measuring the age group of people who have the higher chances of CKD. These measures are analyzed with mean and SD values. Similarly, the parameters are used by the prototype for measuring the performance. For classifying multiple categorizes, it is probable to portray parameters like FP, TP, FN, and TN for every class of $'c'$. Some efficient metrices are used for computing the multi-category classification outcomes. TP rate, accuracy, and F-rank value computation is analyzed for all categories and precision are measured as in Eq. (5.1):

$$Rate - T_p = \frac{T_p}{T_p + F_n} \tag{5.1}$$

This true positive is the measure of total true values attained during classification (true members classified for attaining better outcomes). TP rate is not sufficient for evaluating the performance comprehensively during classification under one category and therefore it is probable to evaluate the class accuracy in Eq. (5.2):

$$Accuacy\ (c) = \frac{T_p}{T_p + F_p} \tag{5.2}$$

F-rank computation is evaluated based on precision validation and considered as harmonic ways for recall and accuracy. This is expressed as in Eq. (5.3):

$$f = 2 * \left(\frac{Accuracy * recall}{Accuracy + recall}\right) \tag{5.3}$$

Similarly, F-computation is evaluated for $'c'$ class is expressed as in Eq. (5.4):

$$f = 2 * \left(\frac{Accuracy\ (c) * T_p rate}{Accuracy\ (c) + T_p\ rate}\right) \tag{5.4}$$

The NB classifier performance while testing the dataset is presented in Table 5.2. From the table $t_l, t_a, and\ t_h$ specifies True Low, Average and High classification notations, while $a_l, a_a\ and\ a_h$ specifies the accuracy as low value , average and higher with classification values respectively. The precision levels based on gain rate, data extraction and Gini index are provided based on CbH using the classifier model as depicted in Table 5.2.

Table 5.2 Naive Bayes Scattering Precision

Category	t_l	t_a	t_h	**Precision**
a_l	589	46	16	92
a_a	31	86	19	87.4
a_h	9	19	79	85.6
Recollection	83	81.2	81	88.8

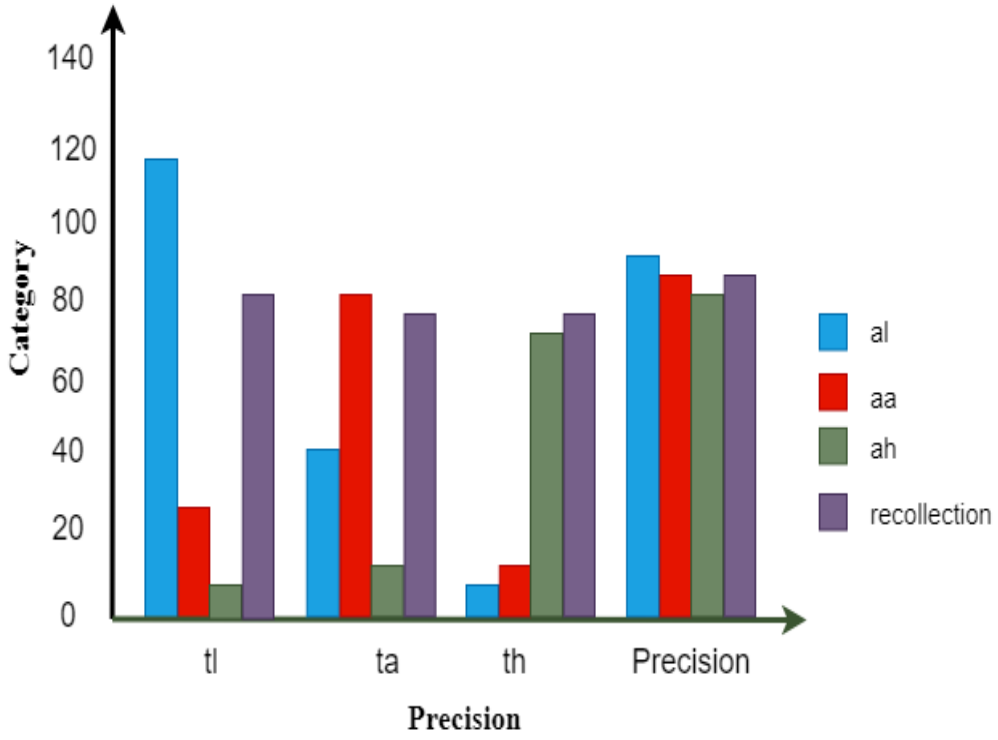

Fig 5.3 NB Scattering Precision

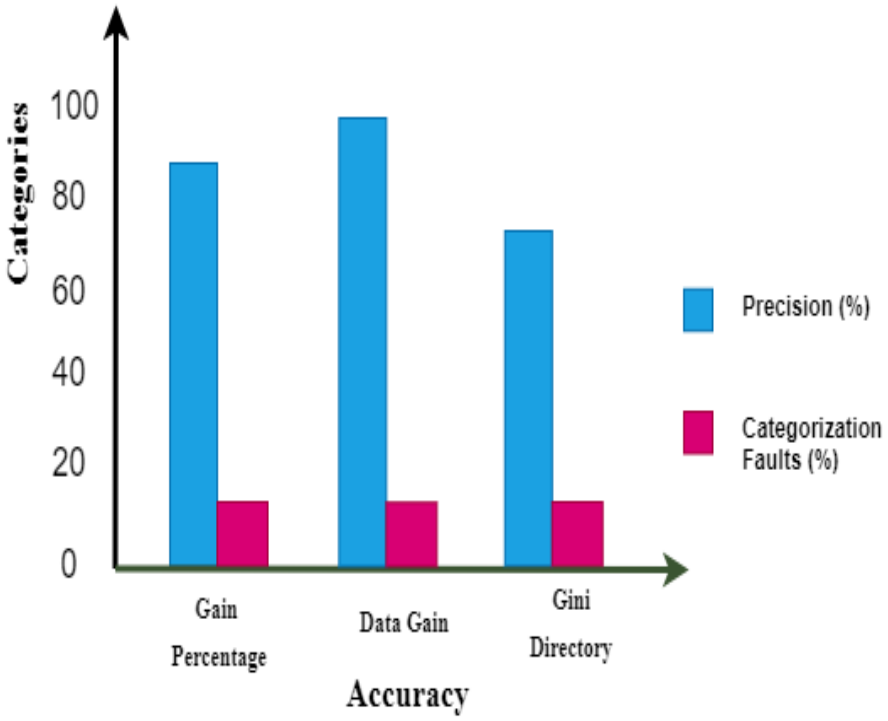

Fig 5.4 NB Scattering Accuracy

The NB classifier performance is analyzed and computed on error/fault rate. The comprehensive analysis is based on fault rate for all the available samples. It is depicted in Fig 5.3 and 5.4 respectively.

Table 5.3 Naive Bayes Scattering Accuracy

Parameters	Precision (%)	Classification errors (%)
Gain Ratio	82.25	10.08
Information Gain	91.5	9.09
Gini Index	72.25	22.35

The precision and classification error based on NB-CbH is portrayed in Table 5.3. Here, parameters like Information Gain, Gain ratio, and Gini Index are considered. Precision is provided in percentage (%) where the values are 82.25%, 91.5% and 72.25% respectively. Similarly, the classification error values are 10.08, 9.09, and 22.32 respectively.

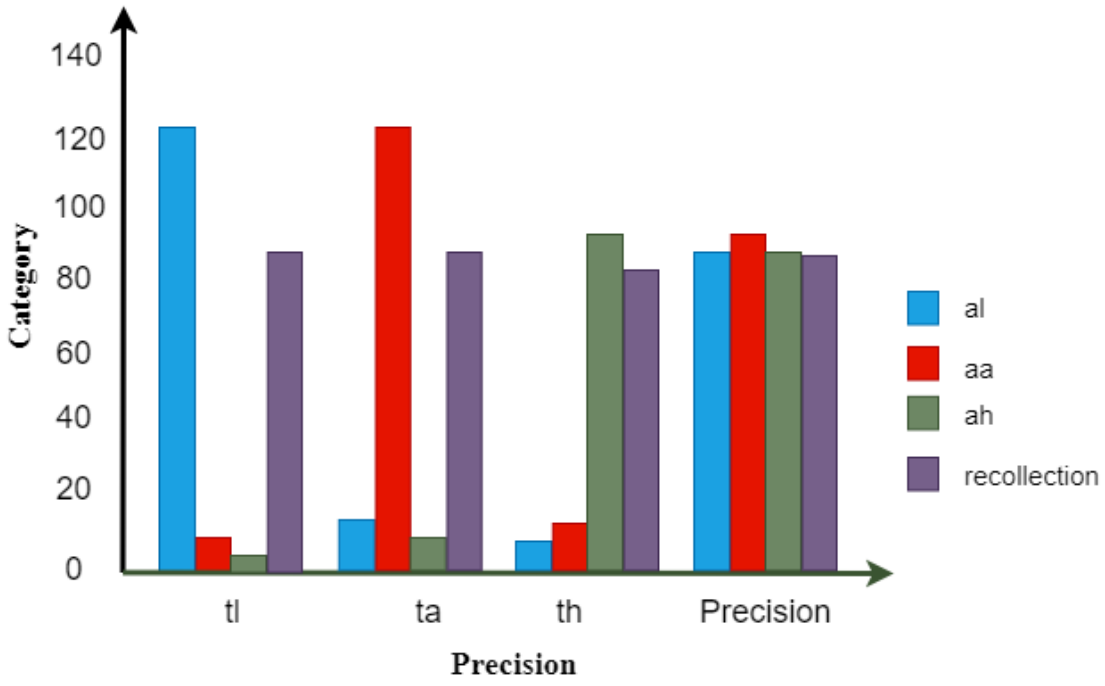

Fig 5.5 Performance of CbH with Precision

Table 5.4 Performance of CbH with Precision

Category	t_l	t_a	t_h	Precision
a_l	626	16	8	90.1
a_a	11	121	16	91.3
a_h	3	9	91	92
Recollection	86	83	81	90.3

Table 5.4 depicts the performance of CbH for measuring precision with average, low and higher values. The average precision value is 90.2% which is higher than other models for predicting the CKD in earlier stage. Fig 5.5 shows the graphical representation of CbH model for precision evaluation.

LDL> 120.450

VLDL>29.564

LDL>154.520

VLDL>43.012; high {low=1, average = 0, high = 61}

VLDL 43.012

Age >40: high {low = 0, average=1, high=21}

Age 40

Age >39.5: average {low = 0, average = 3, high = 0}

Age 39.5: high {low = 0. Average =0, high = 2}

LDL 152.520

BP = 120/80

VLDL > 43.012

Fasting 133.0: high {low = 0, average = 1, high = 0}

Fasting 133.0: medium {low = 0, average = 7, high = 0}

VLDL > 59.012: average {low = 0, average = 24, high = 0}

BP = 130/80

VLDL >58.658: high {low = 0, average = 0, high = 0}

VLDL 59.012

Table 5.5 Precision of Naive Bayes classifier and Choice Based Hierarchy Classifier Scheme

Scheme	Precision
NB	88.6
CbH	90.3

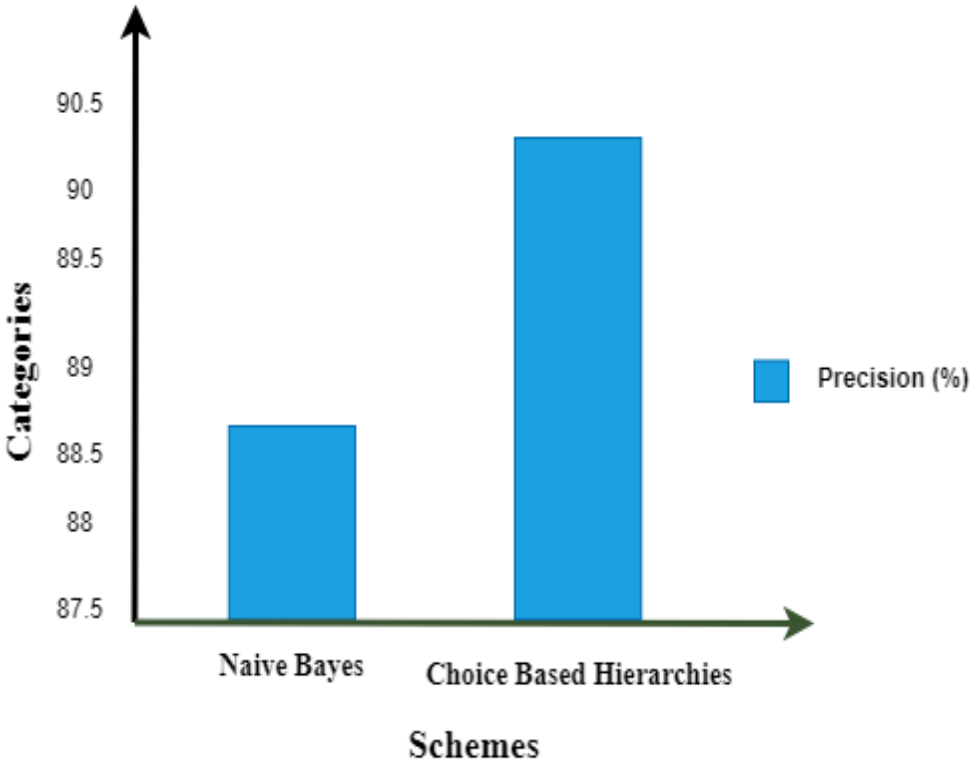

Fig 5.6 Precision of NB and CbH Classification Scheme

Table 5.5 shows the precision values of NB classifier and CbH approaches. NB shows 88.7% and CbH shows 90.2% precision. Fig 5.6 gives the graphical representation of precision comparison of NB and CbH models. From this analysis, it is known that NB classifier and CbH performances better in predicting the features that are related to CKD. With the identification of features in earlier stage leads to better treatment of CKD. Thereby, the quality of human lives can be improved. However, there are some shortcomings that are related to NB classifier. To overcome

this issue, ANFIS model is used for disease prediction. The numerical results attained from ANFIS are discussed in the section given below.

5.3. Performance evaluation of ANFIS model

With ANFIS prediction model, the prediction is done with the evaluation of metrics like precision. The results of ANFIS models is disclosed to analyze the threat levels of CKD. The feature extraction is based on the tree-based grouping/clustering. Here, RF model is used for feature selection and ANFIS is applied for classification. Three different sets are generated to locate the position of individual clusters to evaluate the threats of other related features. Group 1 are unique or identical with improved HB level, raised sugar levels, packet cell, and minimal blood urea, red blood cells with cell clumps and without DM. These people are comes under group 1 and they possess higher chance of CKD.

Most people comes under group 2; but, however lesser features than group 1. However, sugar level is measured to be 1.01 that creates group 2. This category is also supposed to have higher chances of CKD. People with lesser serum creatinine also come under this environment. It is observed with the attained values, some features like diabetes mellitus, hyper-tension and raised sugar levels provides the higher threat level of gaining kidney/renal failures.

a. Confusion Matrix

Confusion matrix for ANFIS model by applying the rule set over the dataset and the outcomes for classifying the threats are specified in Table 5.6. The basic confusion matrix is given in Fig 5.7.

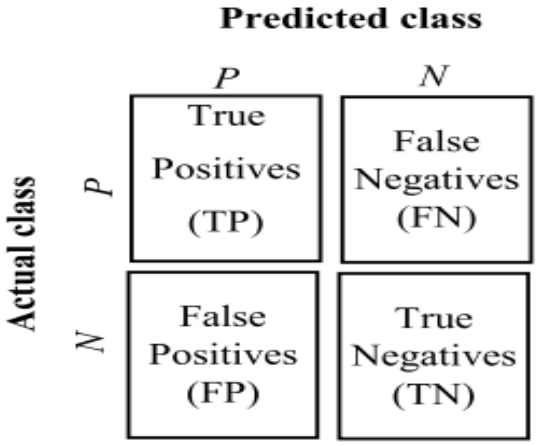

Fig 5.7 Basic confusion matrix

Table 5.6 Confusion Matrix

Original	Forecast	
	1	2
1	12	0
2	0	42

Table 5.7 Precision Estimation for ANFIS

Precision	100%
Sensitivity	100%
Specificity	97%

Table 5.8 Comparison of performance measures

Metrics	NB	ANFIS	Choice based hierarchies
Precision	88.7%	100%	90.2%
Sensitivity	84%	100%	94%
Specificity	82%	97%	92%

The rate of classification error is 0. The precision of ANFIS model is depicted as in Table 5.7. Here, metrics like precision, sensitivity, and specificity is evaluated. The precision and sensitivity is 100%; while specificity is 97%. The anticipated model gives better results in CKD prediction in earlier stage and assists the nephrologists to give proper treatment during the time of critical issues. Table 5.8 depicts the comparison of performance metrics like precision, sensitivity, specificity of NB, ANFIS, and choice based hierarchies respectively.

Fig 5.8 Comparison of performance metrics

Fig 5.8 depicts that comparison of performance metrics where the performance of ANFIS model is higher than NB and CbH respectively. The precision value of NB and CbH is 11.3% and 9.8% lesser than ANFIS model. The ANFIS model shows 100% precision value and outperforms the other approaches. The sensitivity value of NB and CbH is 16% and 6% lesser than ANFIS model. The ANFIS model shows 100% sensitivity value and outperforms the other approaches. Finally, the specificity value of NB and CbH is 15% and 5% lesser than ANFIS model. The ANFIS model shows 97% precision value and outperforms the other approaches. From the above observations, it is known that the performance of ANFIS is higher than conventional approaches like NB and choice based hierarchy model.

5.4. Summary

This chapter discusses about the numerical values attained with the evaluation of proposed NB-CbH and ANFIS model. Initially, feature selection is performed which is followed by

classification approach. The precision of NB-CbH is 90.2%, and ANFIS model is 100% respectively. The proposed model is helpful in predicting CKD in earlier stage. This predictor model helps the nephrologists in taking appropriate decision during the critical time. This can helps for highlighting the quality of life over the patients worldwide. However, in order to enhance the accuracy and classification rate of the predictor model, deep learning approaches can be applied in the future.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH DIRECTION

The target objective of this work is to model an effectual CDSS for CKD predicting using ML approaches. The preliminary goal is to design a predictor model that concentrates on effectual disease prediction in earlier stage. The successive objective is model an effectual classifier to enhance the classification/prediction accuracy with reduced number of CKD features. The selection of feature information is based on the extensive analysis with the literature for designing a predictor model for medical applications. The performance of the anticipated model is to test the functionality of the classifier for medical applications. The model performance is done with data collection from online available sources. The simulation is done in WEKA environment that runs of 64 bit Windows Operating System, 2 GHz speed, and Intel 13 processor on 8 GB RAM.

6.1. SUMMARY OF RESEARCH PERFORMED

In this research work, classification plays a dominant role in prediction of CKD for the diabetic patients with features like age, gender, alcohol, smoking, cholesterol HDL and so on. The classification process is performed in two diverse phases: Naive Bayes classifier with choice based Hierarchical process, Neuro-Fuzzy classifier model. The selection of features from dataset is extremely essential as dominant feature plays a crucial role in disease prediction. Here, the metrics like accuracy and precision are provided with huge significance. These classifiers help the healthcare experts to make proper decision during the time of critical conditions. Also, it examines the dominant feature in the initial phase to enhance the quality of patients' life.

All the afore-mentioned approaches focus on modelling a predictor model with Machine Learning approaches that are utilized as an assisting tool for the Nephrologists during Decision Making process. Each approach possesses its own competency to deal with issues of the specific

field and it is done in a supportive manner. The outcomes are determined to be more robust and intelligent by offering appropriate solutions when compared to traditional approaches.

6.2. CONTRIBUTION OF THE DISSERTATION

The significant contributions of this research work are to model an effectual Decision Support System for CKD prediction using Machine Learning approaches. When comparing the performance of the proposed model with prevailing data-driven methods and human experts, the performance of the anticipated model shows superior outcomes with precision and accuracy. In this regard, the significance of the proposed models is explained below:

- Consider an online available diabetic dataset for CKD disease prediction at earlier stages.
- Examine the most dominant features that influences and triggers CKD for the patients.
- Perform classification with Naive Bayes classifier and Neuro-Fuzzy model to predict the occurrence of CKD and Non-CKD respectively.
- Analyze the metrics like accuracy, precision, F-measure, recall, and confusion matrix for measuring the reliability of the anticipated model.

The anticipated model uses diverse techniques to deal with the problems over the learning field in a cooperative manner than the competitive manner. For modelling an effectual Clinical Decision Support System, data classification is considered to be most essential tasks which are essential for in-depth computation of data. With the diverse available approaches in literature, this research work adopts Naive bayes and Neuro-Fuzzy model to construct an effectual classification system as it offers a facility to deal uncertainty, reasoning, human perception, and decision making during complex environments.

6.3. SCOPE FOR FUTURE RESEARCH ENHANCEMENT

Even though the nephrologists are not recognizable with preliminary ideas of AI in current situation, experts are collaborative with AI researchers and nephrologists in future. It is

recommended to apply AI techniques to construct huge database for CKD with enormous samples and to establish an effectual model that should be extensively applied for treatment and diagnosis of CKD. The accuracy and objectives of AI approaches are extensively applied in pathological diagnosis to help in recognizing the variance among the diseases which are not distinguishable to naked eyes. With this condition, various applications are designed to assist patients for predicting the renal biopsy outcomes and identify renal prognosis. With the advancements in data preservation, establishment of resource sharing platforms, and processing technologies, disease risk models dependent on multi-center enormous data is more reliable. When the performance of the CDSS is considered to be higher, it may substitute the renal biopsy results to acquire a non-invasive diagnosis. The future is unpredictable, however everything is considered to be predictable.

LIST OF PUBLICATIONS

1.Sreeji S & Balamurugan B 2020, Paper Published For “A Survey of Big Data Analytics in HealthCare Industry”, International journal of Advanced Science and Technology Volume 29(3), pp.14490-14502.

2. Sreeji S & Balamurugan B 2020, Paper Published For “Utilizing Schemes for Detecting the Threat Levels for Chronic Kidney Disease”, “Materials Today Proceedings, Elsevier. https://doi.org/10.1016/j.matpr.2020.12.138

3. Sreeji S & Balamurugan B 2020, Paper Published for “A Novel Algorithm for Prediction of Chronic Kidney Risks Using Machine Learning Schemes “Materials Today Proceedings, Elsevier.

https://doi.org/10.1016/j.matpr.2020.11.780

International Conference:

1.Paper Presented the title Big Analytics in Healthcare Industry: A Survey in 3rd International Conference on Advanced Scientific Innovation and Science, Engineering and Technology (ICASISET2020) organized by Bharath Institute of Higher Education and Research, Chennai.

REFERENCES

1. A Manish Kumar, ―Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm‖ International Journal of Computer Science and Mobile Computing, Vol. 5, Issue. 2, February 2016.

2. S.Ramya, Dr. N.Radha, ―Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms‖ International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

3. Jyoti Saini, R.C Gangwar and Mohit Marwaha, ―A Novel Detection For Kidney Disease Using Improved Support Vector Machine‖ International Journal of Latest Trends in Engineering and Technology Vol.(7)Issue(4), 2015.

4. Parul Sinha, Poonam Sinha, ― Comparative Study of Choric Kidney Disease Prediction using KNN and SVM‖ , International Journal of Engineering Research and Technology, Vol(4), Issue-12, 2015.

5. Neha Sharma, Er.Rohit Kumar Verma, ― Prediction of Kidney Disease by using Data Mining Tecniques‖ International Journal of Advanced Research in Computer Science and software Engineering, Vol 6, Issue 9, September 2016.

6. Dr. S. Vijiayarani, Mr. S. Dhayanand, ― Kidney Disease Prediction Using Svm And Ann Algorithms ― International Journal of Computing and Business Research(IJCBR), Volume 6, Issue2, March 2015.

7. Sai Presad Potharaju, M.Sreedevi, ― Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data‖ Journal of Engineering Science and Technology Review 9(5) (2016).

8. Lambodar Jena, Narendra Ku. Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease ― International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11).

9. L.Jerlin Rubini, Dr.P.Eswaran, ―Generating comparative analysis of early stage prediction of Chronic Kidney Disease‖ International Journal Of Modern Engineering Research (IJMER) ISSN: 2249–6645 ,Vol. 5 ,Iss. 7 ,2015.

10. Morteza Khavanin Zadeh , Mohammad Rezapour , Mohammad Mehdi Sepehri, ―Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients‖ International Journal of Hospital Research, 2012.

11. Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq, ―Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease‖, International Journal of Database Management Systems (IJDMS), Vol.8, No.3, June 2016.

12. Pushpa M. Patil, ―Review on Prediction of Chronic Kidney Disease using Data Mining Techniques‖, International Journal of Computer Science and Mobile Computing, Vol. 5, ISSN 2320–088X, Issue. 5, May 2016.

13. Mohammad Rezapour, Morteza Khavanin and Mohammed Mehdi Sepehri, ―Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients‖ Computational and Mathematical Methods in Medicine, Volume 3, 2013.

14. R.Sujatha, Dr.Ezhilmaran, ―Performance Analysis Of Data Mining Classification Techniques For Chronic Kidney Disease‖ International Journal Of Pharmacy & Technology, ISSN: 0975-766X, Vol-6, 2016.

15. F Al-Turjman, MH Nawaz, UD Ulusar, Intelligence in the Internet of Medical Things era: A systematic review of current and future trends ,Computer Communications 2019

16. M Bhuvaneswari, GN Balaji, F Al-Turjman, Machine Learning Parameter Estimation in a Smart-City Paradigm for the Medical Field,Smart Cities Performability, Cognition, & Security, 139-151

17. BD Deebak, F Al-Turjman, M Aloqaily, O Alfandi, An Authentic-Based Privacy Preservation Protocol for Smart e-Healthcare Systems in IoT, IEEE Access 7, 135632-135649

18. A Manish Kumar, ―Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm‖ International Journal ofComputer Science and Mobile Computing, Vol. 5, Issue. 2, February 2016.
19. S.Ramya, Dr.N.Radha, ―Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms‖ International Journal of InnovativeResearch in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.
20. Jyoti Saini, R.C Gangwar and Mohit Marwaha, ―A Novel Detection For Kidney Disease Using Improved Supportvector Machine‖ International Journal of Latest Trends in Engineering and Technology Vol.(7)Issue(4), 2015.
21. Parul Sinha, Poonam Sinha, ― Comparative Study of Choric Kidney Disease Prediction using KNN and SVM‖ , International Journal ofEngineering Research and Technology, Vol(4), Issue-12, 2015.
22. Neha Sharma, Er.Rohit Kumar Verma, ― Prediction of Kidney Disease by using Data Mining Tecniques‖ International Journal of AdvancedResearch in Computer Science and software Engineering, Vol 6, Issue 9, September 2016.
23. Dr. S. Vijiayarani, Mr. S. Dhayanand, ― Kidney Disease Prediction Using Svm And Ann Algorithms ― International Journalof Computing and Business Research(IJCBR), Volume 6, Issue2, March 2015.
24. Sai PresadPotharaju, M.Sreedevi, ― Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data‖ Journalof Engineering Science and Technology Review 9(5) (2016).
25. Lambodar Jena, Narendra Ku. Kamila ,‖ Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease ―International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11).

26. L.Jerlin Rubini, Dr.P.Eswaran, ―Generating comparative analysis of early stage prediction of Chronic Kidney Disease‖ International Journal OfModern Engineering Research (IJMER) ISSN: 2249–6645 ,Vol. 5 ,Iss. 7 ,2015.

27. MortezaKhavaninZadeh , Mohammad Rezapour , Mohammad Mehdi Sepehri, ―Data Mining Performance in Identifying the Risk Factors ofEarly Arteriovenous Fistula Failure in Hemodialysis Patients‖ International Journal of Hospital Research, 2012.

28. Basma Boukenze, Hajar Mousannif and AbdelkrimHaqiq, ―Performance of Data Mining Techniques to Predict in Healthcare Case Study:Chronic Kidney Failure Disease‖, International Journal of Database Management Systems (IJDMS), Vol.8, No.3, June 2016.

29. Pushpa M. Patil, ―Review on Prediction of Chronic Kidney Disease using Data Mining Techniques‖, International Journal of Computer Scienceand Mobile Computing, Vol. 5, ISSN 2320–088X, Issue. 5, May 2016.

30. Mohammad Rezapour, MortezaKhavanin and Mohammed Mehdi Sepehri, ―Implementation of Predictive Data Mining Techniques forIdentifying Risk Factors of Early AVF Failure in Hemodialysis Patients‖ Computational and Mathematical Methods in Medicine, Volume 3,2013.

31. R.Sujatha, Dr.Ezhilmaran, ―Performance Analysis Of Data Mining Classification Techniques For Chronic Kidney Disease, International Journal Of Pharmacy & Technology, ISSN: 0975-766X, Vol-6, 2016.

32. Yao D, Yang J, Zhan X. A novel method for disease prediction: hybrid ofrandom forest and multivariate adaptive regression splines. J Comput. 2013;8(1):170–7.

33. R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Machine learning: anartificial intelligence approach. Springer Science & Business Media, 2013.

34. Davis DA, Chawla NV, Christakis NA, Barabási A-L. Time to CARE: acollaborative engine for practical disease prediction. Data Min Knowl Disc.2010;20(3):388–415.

35. McCormick T, Rudin C, Madigan D. A hierarchical model for association rulemining of sequential events: an approach to automated medical symptomprediction; 2011.

36. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models toassess risk of type 2 diabetes, hypertension and comorbidity: machinelearningalgorithms and validation using national health data from Kuwait-acohort study. BMJ Open. 2013;3(5):e002457.

37. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.

38. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013.

39. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.

40. I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, vol. 3, 22, pp. 41–46: IBM New York.

41. Toshniwal D, Goel B, Sharma H. Multistage Classification for Cardiovascular Disease Risk Prediction. In: International Conference on Big Data Analytics; 2015. p. 258–66. Springer.

42. Forssen H, et al. Evaluation of Machine Learning Methods to Predict Coronary Artery Disease Using Metabolomic Data. Stud Health Technol Inform. 2017;235: IOS Press:111–5.

43. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. In: AMIA annual symposium proceedings, vol. 2012; 2012. p. 606. American Medical Informatics Association.

44. Malik S, Khadgawat R, Anand S, Gupta S. Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. SpringerPlus. 2016;5(1):701.

45. Lundin M, Lundin J, Burke H, Toikkanen S, Pylkkänen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology. 1999;57(4):281–6.

46. Aneja S, Lal S. Effective asthma disease prediction using naive Bayes—Neural network fusion technique. In: International Conference on Parallel, Distributed and Grid Computing (PDGC); 2014. p. 137–40. IEEE.

47. Mansoor H, Elgendy IY, Segal R, Bavry AA, Bian J. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach. Heart Lung. 2017;46(6):405–11.

48. Kim J, Lee J, Lee Y. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthc Inform Res. 2015;21(3):167–74.

49. Thenmozhi K, Deepika P. Heart disease prediction using classification withdifferent decision tree techniques. Int J Eng Res Gen Sci. 2014;2(6):6–11.

50. Marikani T, Shyamala K. Prediction of heart disease using supervised learning algorithms. Int J Comput Appl. 2017;165(5):41–4.

51. Lu P, et al. Research on improved depth belief network-based prediction of cardiovascular diseases. J Healthc Eng. 2018;2018:1–9.

52. Khateeb N, Usman M. Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. In: Proceedings of the International Conference on Big Data and Internet of Thing; 2017. p. 21–6. ACM.

53. Venkatalakshmi B, Shivsankar M. Heart disease diagnosis using predictive data mining. Int J Innovative Res Sci Eng Technol. 2014;3(3):1873–7.

54. Lynch CM, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform. 2017;108:1–8.

55. Eskidere Ö, Ertaş F, Hanilçi C. A comparison of regression methods forremote tracking of Parkinson's disease progression. Expert Syst Appl. 2012;39(5):5523–8.

56. Avci E, Extraction AD (2018) Performance Comparison of Some Classifiers on Chronic Kidney DiseaseData.

57. Bommanna Raja K, Madheswaran M (2007) Determination of kidney area independent unconstrained features for automated diagnosis and classification. Int Conf Intell Adv Syst ICIAS 2007:724–729, 2007

58. Chen W, Gou S, Wang X, Li X, Jiao L (2018) Classification of PolSAR images using multilayer autoencoders and a self-paced learning approach. Remote Sens 10(1)

59. Chronic_Kidney_DiseaseDataset. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, Accessedon date: 20-01-2019

60. Chetty SDS, Naganna KSV (2015) Role of Attributes Selection in Classification of Chronic Kidney Disease Patients. Comput. Commun. Secur. (ICCCS), 2015 Int. Conf. on. IEEE, pp. 1–6

61. Dulhare UN (2016) Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier. pp. 4

62. Kannadasan K, Edla DR, Kuppili V (2018) Type 2 diabetes data classification using stacked autoencoders in deep neural networks, pp. 2–7

63. Khamparia A, Pandey B (2019) A novel integrated principal component analysis and support vector machines-based diagnostic system for detection of chronic kidney disease. International Journal of Data Analysis Techniques and Strategies 12(2):1–15

64. Khamparia A, Singh A, Anand D, Gupta D, Khanna A, Arun Kumar N, Tan J A novel deep learning-based multi-model ensemble method for the prediction of neuromuscular disorders. Neural Comput & Applic:1– 13.

65. Kunwar V, Chandel K, Sabitha AS, Bansal A (2016) Chronic kidney disease analysis using data mining classification. Cloud Syst. Big Data Eng. (Confluence), 2016 6th Int. Conf. IEEE, pp. 300–305

66. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26

67. Pujari RM, Hajare MVD (2014) Analysis of Ultrasound Images for Identification of Chronic Kidney. First Int Conf Networks Soft Comput:380–383

68. Qian S, Liu H, Liu C, Wu S, Wong HS (2018) Adaptive activation functions in convolutional neural networks. Neurocomputing 272:204–212

69. Wibawa MS, Maysanjaya IMD, Putra IMAW(2017) Boosted classifier and features selection for enhancing chronic kidney disease diagnose. 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017

70. Adam T, Hashim U (2012) Designing an Artificial Neural Network Model for the Prediction of Kidney problems symptom through the patient ' s metal behavior for pre-clinical medical diagnostic, pp. 27–28

71. Arasu SD, Thirumalaiselvi R (2017) A novel imputation method for effective prediction of coronary Kidney disease. Proc. 2017 2nd Int. Conf. Comput. Commun. Technol. ICCCT 2017, pp. 127–136

72. S. Ahmed, T. Kabir, N. T. Mahmood, and R. M. Rahman, “Diagnosis of Kidney Disease Using Fuzzy Expert System,” 2014.

73. Adem K, Kiliçarslan S, Cömert O (2019) Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder. Expert Syst Appl 115:557–564

74. W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahandawaarachchi,“Performance evaluation on machine learning classificationtechniques for disease classification and forecasting through data analyticsfor chronic kidney disease (CKD),” in Proc. IEEE 17th Int. Conf. Bioinformaticsand Bioengineering, Oct. 2017, pp. 291-296.

75. D. Dua and C. Graff, “UCI Machine Learning Repository,” Irvine, Universityof California, School of Information and Computer Sciences, 2017.[Online]. Available: http://archive.ics.uci.edu/ml.

76. H. Jin, S. Kim, and J. Kim, ``Decision factors on effective liver patientdata prediction," Int. J. Bio-Sci. Bio-Technol., vol. 6, no. 4, pp. 167_178,Aug. 2014.

77. V. Giannouli and N. Syrmos, ``Attitudes of younger and older adults towards kidney diseases in Greece," Health Psychol. Res., vol. 7, no. 2, p. 8230, 2019.

78. E. K. Hashi, M. S. U. Zaman, and M. R. Hasan, ``An expert clinicaldecision support system to predict disease using classi_cation techniques,"in Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE), Feb. 2017,pp. 396_400.

79. U. R. Acharya, H. Fujita, V. K. Sudarshan, M. R. K. Mookiah, J. E. Koh,J. H. Tan, Y. Hagiwara, C. K. Chua, S. P. Junnarkar, A. Vijayananthan,and K. H. Ng, ``An integrated index for identification of fattyliver disease using radon transform and discrete cosine transform featuresin ultrasound images," Inf. Fusion, vol. 31, pp. 43_53, Sep. 2016

80. K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doeleman, andW. T. Freeman, ``Computational imaging for VLBI image reconstruction,"in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016,pp. 913_922.