IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 8, Issue 1, Feb - March, 2020
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

A REVIEW: MACHINE LEARNING BASED CHRONIC KIDNEY DISEASE PREDICTION

Jayashree M[1], Dr. Anitha N[2]
[1]Information Science,VTU,Bangalore,India.
[2]Information science, VTU, Bangalore, India.

Abstract

Long-term Renal infection is a serious health issue that gradually reduces the kidney function. Initial stage of prediction can help, perhaps to lessen the continuous infection of this chronic disease. The exact implication of Long term Renal disease (CKD) or final stage renal disease (FSRD) cannot be estimated accurately. the Symptoms of CKD is not observed properly at the early stage and it becomes difficult to predict,which leads to the permanent damage to the functioning of the Kidney. In Health care it has become a challenging task to predict at the earliest stage by observing the related symptoms and do the necessary test at the earliest. The early prediction of CKD has been deliberately studied using Machine Learning algorithms. different classifier algorithms such as Naïve Bayes, ANN, Ada Boost, Gradient Boost, Random Forest, NN-CBR studied and their performance is being Compared with respect to the accuracy, sensitivity, specificity. Here we will have a brief discussion of these algorithms and their shortcomings with the performance measures with each other. In this Survey, the different Machine Learning methods are reviewed to determine, which algorithm can accurately detect the early stage of the renal problem.

KeyWords: Long term Renal (Chronic Kidney disease),Random Forest,Naïve Bayes, NN-CBR, KNN, Ada Boost and Gradient Boost, Artificial Neural Network

I.Introduction

Long term or Chronic kidney disease is one of the a serious problem observed in adults and older people, if not detected and treated at an early stage it leads to failure of organ and the death rate is increased. Symptoms like blood pressure, diabetes, back pain, anemia in renal area, Darkness or redness in the urine etc. can be observed. By training the patient data and his historical data Using Machine learning algorithm can help the physicians to predict the early stage of CKD. A leading data analytics company which is according to the globalization of data (2019) has specified that CKD is having a poor diagnosis rate unless there are effective ways to be used to diagnose disease at initial stage. The learning model is used to detect the progression of the kidney functioning disease, which helps the physicians to test the patient with the medical data to know which type of patient to treat so that the mortality rate can be decreased.[7].

The individual who is infected with Kidney problem shows the following symptoms like The anemia, weak bones, high Blood Pressure, fatigue, poor nutrition health and nerve damage, Decreased immune response. As the disease progresses it goes to the advanced stages leading to the accumulation of fluids, electrolytes, and wastes in the blood and body [8], if it is not detected at an early stage.

The main excretory organ in the human body is kidney, which removes all the waste materials and toxins from the body. There are approximately 1 million cases of Chronic Renal Disease per year in India. CKD is the one which loses its functionality gradually. if toxins are not removed either in the form of sweat or urine it gets accumulated in the blood with lack of oxygen supply to the body a leading to the death, hence it becomes important to detect the failure at the early stage. It includes balance of water and mineral maintenance, erythropoietin for stimulating RBC production as the main function of Healthy kidneys[9].

CLASSIFICATION OF KIDNEY DISEASE

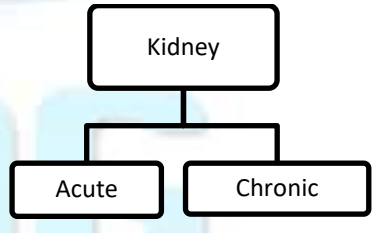

Fig 1: Classification of Kidney disease

Acute Kidney Disease (AKD): The three main causes of AKD includes Lack of blood flow to the kidneys and taking some medicine, which may injure the kidney for a short period of time.

Chronic Kidney Diseases(CKD): Non functioning of the kidney for more than 3 months is categorized as CKD. As symptoms will not be noticed till the major damage to the kidney, detection at the early stage is the first step in treating CKD by observing the major symptoms related to the functionality of the kidney. The CKD infection may increase if the patient has certain long term High blood pressure, Diabetes and some heart related disease and also some genetic diseases which are carried out with family history. The symptoms that need

to be monitored are short growth, Lots of urine or no urine, Anemia, Weak bones, causing swelling in ankles, vomiting, weakness, poor sleep and shortness of breath, Headache, Pale skin, Blood Pressure, Repeated urinary tract infections

II. Related work

Asif slicing [1] The author used a novel approach to classify the CKD using machine learning algorithms. conducted tests on a dataset containing 400 records and 25 attributes which classified the patients having CKD or not CKD and used KNN, random forest and neural network to get results. For feature reduction wrapper method was used which detected CKD with high accuracy.

Gunarathne et.al. [2] compared results of different models on CKD, and concluded that the more accuracy for dataset having 14 attributes achieved using Multiclass Decision forest algorithm gives.

Sahil Sharma, [3]used different classification algorithm on 12 attributes with dataset having 400 records and 24 attributes. calculated results were compared with actual results for calculating the accuracy of prediction results. They used performance metrics like sensitivity, accuracy, precision and specificity. Concluded that decision tree achieves 98.6% accuracy.

S. Ramya and Dr.N.Radha [4] had worked on diagnosis time and improvement of diagnostic accuracy, using different classification algorithms of machine learning. Analysis of various algorithms like BPNN , RBF and RF was done and the analysis results indicate that RBF algorithm gives better results compared to other classifiers and produces 85.3% accuracy.

S. Dilli Arasu [5] has worked on missing values in a dataset of chronic Kidney Disease. the accuracy and prediction results are reduced if there are Missing values in the dataset, therefore Missing values were replaced with recalculated values

Pinar Yildirim [6] The medical decision on CKD is done using neural network algorithm. the effect on preceptor can be measured using the learning rate which is an activation parameter. the sampling algorithms were used.

Vijayarani, S [10]. The prediction of CKD is implemented using ANN and SVM. The system significantly reduces the computation time and also increases system accuracy.

Lambda Jena [11]. The various machine algorithms are used to predict the CKD by considering UCI dataset. The algorithm tested are Multilayer Perceptron (MP), Naive Bayes, But MP algorithm gives better accuracy with good classification results.

Ray Key Chiu [15]. The prediction of CKD is implemented in three different neural networks based on modular neural network, generalized Forward network and back propagation neural network. These models based neural network depending on the genetic algorithm with neural factor. 85% accuracy is achieved with these three models.but highest accuracy produced by back propagation.

U. N. Dulhare.et al[17] Based on stages the action rules are extracted and CKD is detected by using NB with Single R attribute selector.

H.Zhang. et al [18] The early detection of CKD was done with Artificial Neural Network (ANN) models, and continuous degradation of the disease can be reduced with proper diet and treatment.

Aljaaf [19] suggested that predictive analytics can be used as an intelligent solution for early prediction of the disease. the ensemble techniques like AdaBoost and LogitBoost are generally use, and their results are interpreted..

Arif-UL-Islam. et al [20] Detection of CKD was analyzed with the help of boosting algorithms and the rules are derived illustrating the relationship between the various attributes of CKD. The paper used Ant-Miner machine learning algorithm derive rules.

Jerlin Rubini et al [22] The paper used radial basis function network, and logistic regression. multilayer perceptron

Shital Shah et al [23] detecting the CKD using the data mining approach was suggested.

Paul Sinha et al [24] The prediction chronic kidney disease was carried out using decision support system. Comparison of SVM and KNN classifier done on the basis of its accuracy. It has been concluded that KNN is better than SVM classifier

Gabriel R[25].CKD was detected using NN-CBR twin system. The explanation based instance is considered. Here the accuracy of the NN classification is compared. CBR System is evaluated based on the explanation given to the instances.

III. Attributes surveyed for Early Detection of CKD

Initially prediction was done by considering 24 attributes: Blood Pressure, Albumin ,Sugar, Red Blood Cells,Specific Gravity, Pus Cell, Bacteria, Pus Cell clump, Blood Urea, Blood Glucose , ,Serum Creatinin, Sodium, Potassium, Hemoglobin, Age ,Packed Cell Volume ,Count ,Red Blood Cell Count, White Blood Cell, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia ,here 11 numeric 13 nominal were considered. random forest classifier was used with 24 attributes and achieved a0.993 accuracy [1].

Wrapper method used with 12 attributes :Albumin, Specific Gravity ,Red Blood Cells, Serum creatinin, Pus Cell clumps, Sodium, Hemoglobin, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, which detected CKD with maximum Accuracy, to rank the attributes according to their predictive capability it adopted LASSO Regularization method.

To Detect CKD with accurately the classifiers was used considering only 5 attributes: specific gravity, albumin, diabetes mellitus, hypertension and hemoglobin as features. here the major factor serum creatinin was not considered which is critical factor for measuring the CKD.

The 5 attributes: serum creatinin, Blood pressure ,packed cell volume and anemia hypertension were considered to calculate the L-factor[12] and classified as CKD and non-CKD patients based on the L-factor value. According to their evaluation CKD cannot be detected based on their L-factor classifiers. The machine learning [13][14].algorithms such as back propagation neural networks, radial basis functions, random forests and SVM and achieved up to 85.3% accuracy on identifying CKD. Also, goals of study are to comprehensively explore parameters which are related to kidney disease and to introduce a cost effective machine learning approach to detect early CKD instead of the GFR estimation equations[14]

IV. Machine Learning Approaches

A. **Supervised Learning**:A learning algorithm which takes labeled set of input and output data trains a model to generate predictions for the test data is called supervised learning algorithm.

Classification algorithms

a. Support Vector Machine(SVM)

The principal of minimization and statistical learning theory based classifier algorithm is SVM. The efficient separation of classes is determined with help of the hyperplanes (decision boundaries).The data is divided into two classes based on the margin kept on either side of the hyperplane. By increasing the margin on either side of the hyperplane the object decrease the upper bound and generates the error leading to the maximum distance among the hyperplane. SVM works on the data which is linearly separable and linearly Non-separable. only one hyperplane is used for separating the data in case of linearly separable ,the latter uses more than one hyperplane. SVM is less likely to be affected for overfitting of the data than others. SVM are less prone to overfitting of the data [21]. They are extremely accurate for modeling complex nonlinear decision boundaries.

b.*K-Nearest Neighbors(KNN):* The simplest classifier algorithm is k-nearest neighbors. The like samples which fall to near proximity is considered in KNN. For classifying the test data the data which are nearest to training data will be used. In KNN the learning is lazier compared to other learning algorithms. The distribution of data is not assumed and hence follows non parametric approach. The calculation of distance between the test and training data is done using the Euclidian distance. Classification of test data is decided based on the training data which are belong to K-nearest neighbors. The value of K is considered small integer. It becomes difficult to classify the classes when the value of K increases. The best value of K is chosen based on Cross-validation approach.

c.Artificial Neural Network(ANN): Artificial neural network (ANN) is similar with respect to the behavior of the biological network of neurons. ANN is a tool used for modeling data which performs multiple parallel computations and hence used for representing and simulating the relationships between inputs and outputs by. The important components of ANN are input layer hidden layer and the output layer. Here the model is trained to learn by adjusting the weights between the three layers ,i.e output to hidden and hidden to input. The neuron weighted inputs are converted to output using activation function.. The Sigmoid function is used for non linear data.

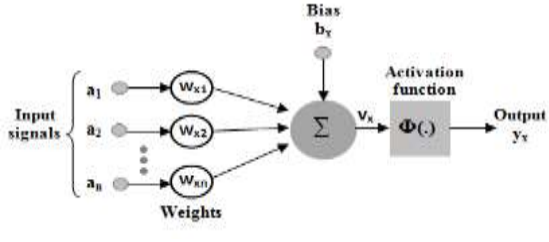

Fig.2: Neural Network Model

d.Extreme Learning Machine (ELM): ELM is a feed forward network having single hidden layer (SLFNs). Chronic Kidney Disease prediction can be done using kernel-based ELM. The ELM can be used as both regressor and classifier[31].the hidden layers are projected by ELM randomly. analytical calculation is done using the method of least square in order to get the output weight. The repetitive training of ELM is not required. Hence it is faster than traditional SLFNs. The kernel functions present in ELM are the radial basis function, linear function [19] and polynomial function. the learning of Extreme Learning Machine (ELM) is based on SLFNs. The fine tuning of the hidden weights is not required in ELM, instead it the hidden nodes are chosen randomly and determines output weight using an analytical calculation.ELM has good performance and extremely faster than the standard feed forward learning. The different kernel functions considered based on analytical calculation are:

i. Linear kernel
The <x,z> are given as internal variables and their values are added to c which is a constant choosen arbitrarily. Kernel algorithms is given by $K(x,z) = x^Tz + c$. Where K is matrix of the kernel, C is regularization parameter, and W is the weighted matrix

ii.Polynomial Kernel
The normalized training data uses Polynomial kernels [25].This kernel uses the constant c ,slope alpha, and the polynomial degree d as modifiable parameters, Given by $K(x,z) = (ax^Tz + 1)^d$

iii.Radial Basis Function Kernel.
RBF kernel uses initialization of centers and dimensions randomly, analytical calculation is used for RBF kernel network output weights. The best parameters were selected using the given equation which gives the most accuracy [26](x,z)=exp $(-\gamma\|x-z\|^2)$.

B. Un supervised learning: Clustering

The set of data having the similar patterns or clusters are identified and grouped to known set of results, the results are analyzed and can be used as classification technique .It needs to accurately find the centroids for which clustering algorithms are used.

a.K-Means Clustering Algorithm

one of the Unsupervised Learning algorithm is K-Means, The data set has to be classified based on the number of clusters, or centroids(K) in which,the model is based upon centroid clustering. The location is optimized by the series calculation of centroids. A mapping of each point is done to distinct cluster for which the distance is minimum. set of observations given are $(x_1,x_2, ..., x_n)$,where each observation is a d-dimensional real vector, the K-Means clustering aims to partition the n observations into $k(\leq n)$ sets $S = S_1,S_2,...,S_k$ to minimize the within-cluster sum of squares (WCSS). Optimize function is given by

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Here the training set is preprocessed using the L-factor calculation by considering the major factors like Sugar Levels,Blood Pressure, and Anemia with unusual Creatinin levels.

C.**Boosting algorithms**

Many weak classifiers are pooled to form strong classifier in order to improve the classification accuracy. Boosting algorithms considered are:

Classifier	Sensitivity	Accuracy	Specificity
SVM	**1**	.7375	0.7600
KNN	0.9920	.7875	0.7600
ANN	0.9744	.80	0.7107
K-Mean	86.7	88	89.86
CBR	84	86	88
Adaboost	97.2	97.33	97
XGBoost	0.96	0.97	0.97

Ada Boost is an assembling learning algorithm for data classification. It collects the data with equal distribution and finds the classifier which has a good weight below the threshold. The Algorithm updates the weights and focuses on erroneously classified samples. It performs certain Iterations to improve the performance of the classification.

XGBoost: XGBoost supports both regression and classification. Boosting is an ensemble technique which was developed to improve the performance and computational speed. If existing models generate the error the newer models are added to adjust the error. the Models are added recursively to find the improvements and it is stopped once no more

improvement is found. A gradient descent algorithm is used to minimize the loss.

V. Steps in Processing of data sets

The input data set is collected from UCI Repository. Preprocessing of the data is carried out to remove the noise, to fill the missing data in the datasets. the performance measures can be evaluated using Machine Learning Algorithms by detecting the disease at the initial stage. damage to the kidney occurs slowly if the symptoms are not noticed at the early stage. The factors causing like High blood pressure, Diabetes and cardiovascular disease are also consider for kidney infection apart from these factors the history of family background health states is also included.

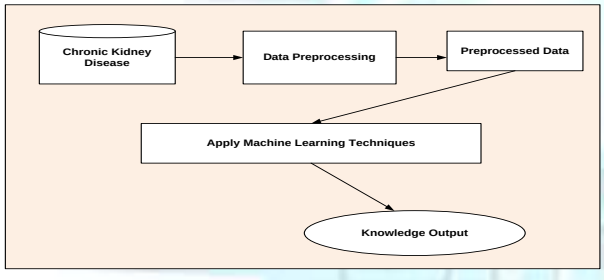

Fig 3: Processing of Kidney Data set

Preprocessing of Data is the basic operation which include cleaning and reduction of data. Cleaning of data process includes noisy data removal and missing data field. the unnecessary dimensionality reduction in data is done by removal process. the effective number of variables can be considered using this process. suitable dataset is selected by the target dataset selection process. Finally the data samples which are relevant to the analysis task are retrieved from the database.

VI. Comparison of the results

The ratio of summation of true positive and true negative to the total number instances gives the accuracy.
sensitivity =TP/TP + FN, specificity =TN/TN + FP

Table 1:Comparison of metrics

VII. CONCLUSION

This survey paper has studied and analyzed the various Machine Learning (ML) algorithm which are helpful for early detection of CKD. From the study of different ML algorithm it has been observed that accuracy of the system is varied for different algorithms and decides the efficient techniques and tools to be used for implementing the system. The accuracy can be predicted from various machine learning algorithms like KNN, Neural Network, Support Vector Machine etc. ML algorithms produced good prediction results if any appropriate techniques are applied to input datasets. Further study of this survey paper needs to strengthen the available data set must be higher because of increasing the size of the data set and considering the missing attribute parameters will help to improve the system performance. To initiate the machine learning model by using significant feature will help to predict CKD with best accuracy and will need higher recorded data set with less missing values been applied for CKD classification.

References

[1]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[2]Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)",2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[3]. Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," July18, 2016.

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

[5]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining Techniques",International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505.

[6]. L. Rubini, "Early stage of chronic kidney disease UCI machine learning repository,"2015. [Online].Available:http://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.

[7].Sunethra Kanthi Gunatilake1, S. Sunil Samaratunga2 and Ruvini Takshala Rubasinghe1,Chronic Kidney Disease (CKD) in Sri Lanka - Current Research Evidence Justification: A Review, Sabaragamuwa University Journal Volume 13 Number 2; December 2014.*

[8].S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning : a review," IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[9].Kidney Failure (ESRD) Causes, Symptoms, & Treatments .www.kidneyfund.org › kidney-disease › kidney-failure

[10].Vijayarani, S and S.Dhayanand, Kidney disease Prediction Using SVM and ANN Algorithms, International Journal of Computing and Business Research, 6(2), 2015.

[11] Lambodar Jena and Narendra Ku.Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease. International Journal of Emerging Research in Management &Technology, 4(11), pp.110-118, 2015.

[12] A. Dubey, "A classification of ckd cases using multivariate k-means clustering." international Journal of Scientific and Research Publications (IJSRP), vol. 5, August 2015.

[13] P. Sinha and P. Sinha, "Comparative study of chronic kidney disease prediction using knn and svm," International Journal of Engineering Research and Technology, vol. 4, no. 12, 2015.

[14] S. Ramya and N. Radha, "Diagnosis of chronic kidney disease using machine learning algorithms," International Journal of Innovative Researchin Computer and Communication Engineering, vol. 4, no. 1, 2016.

[15]Ruey Kei Chiu and Renee Yu-Jing, Constructing Models for Chronic Kidney Disease Detection and Risk Estimation. Proceedings of 22nd IEEE International Symposium on Intelligent Control, Singapore, pp. 166-171, 2011.

[16]. Jiawei han and Micheline Kamber, (2006) "Data mining Concepts and Techniques,Second edition", Morgan Kaufmann publishers, by Elsevier Inc. All rights reserved.

[17] M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 300-303.

[18] H. A. Wibaa, I. Malik and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," 2018 2nd International Conference on Informaticsand Computational Sciences (ICICoS), Semarang, Indonesia, 2018, pp. 1-4

[19]U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[20]H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 1351-1356.

[21]. Sunitha Devi1. P, Sowjanya. CH, Sunitha. K.V.N, (2014) "A Review of Supervised Learning Based Classification for Text to Speech System", International Journal of Application or Innovation in Engineering & Management, Volume 3, Issue 6, June page no 79-86

[22]Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar,Bangladesh, 2019, pp. 1-6.

[23]R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naïve Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 679-684.

[24]Jerlin Rubini and Eswaran," Generating Comparative Analysis of Early Stage Prediction of Chronic Kidney Disease", International Journal of Modern Engineering Research, 2015, Vol. 5, No. 7, pp. 49-55.

[25]Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms: A Survey",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8

[26]Parul Sinha and Poonam Sinha, "Comparative study of chronic kidney disease prediction using KNN and SVM", International Journal of Engineering Research & Technology (IJERT), December 2015 (ISSN: 2278-0181), Vol. 4, Issue 12, PP. 608-612,

[27]N.Pavithra, Dr. R. Shanmugavadivu "Efficient Early Risk Factor Analysis of Kidney Disorder Using Data mining Technique", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017,pp.1690

[28] Gabriel R. Vasquez-Morales "Explainable Prediction of Chronic Renal Diseasein the Colombian Population Using Neural Networks and Case-Based Reasoning" Digital Object Identifier 10.1109/IEEE ACCESS.2019.2948430,

[29]Jiongming Qin Lin Chen "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease" DOI 10.1109/ACCESS.2019.2963053, IEEE Access 2017.

[30]Nikita, Balwinder Kaur, Dr. Harsh Sadawarti, Dr. Jimmy Singla "A Neuro-Fuzzy Based Intelligent System For Diagnosis Of Renal Cancer" INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 01 , JANUARY 2020 ISSN 2277-8616.

[31] K. Anam and A. Al-Jumaily, "A novel extreme learning machine for dimensionality reduction on finger movement classification using sEMG," Int. IEEE/EMBS Conf. Neural Eng. NER, vol. 2015–July, pp. 824–827, 2015.

[32] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," AI communications, vol. 7, pp. 39-59, 1994.

[33] M. T. Keane and E. M. Kenny, "How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBRtwin-systems," in International Conference on Case-Based Reasoning, 2019