International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Machine Learning Models and Neural Network Techniques for Predicting Uddanam CKD

K. B. Anusha, T. PanduRanga Vital, K. Sangeeta

Abstract*: Chronic kidney disease (CKD) is one of the most widely spread diseases across the world. Mysteriously some of the areas in the world like Srilanka, Nicrgua and Uddanam (India), this disease affect more and it is cause of thousands of deaths particular areas. Now days, the prevention with utilizing statistical analysis and early detection of CKD with utilizing Machine Learning (ML) and Neural Networks (NNs) are the most important topics. In this research work, we collected the data form Uddanam (costal area of srikakulam district, A.P, India) about patient's clinical data, living styles (Habits and culture) and environmental conditions (water, land and etc.) data from 2016 to 2019. In this paper, we conduct the statistical analysis, Machine Learning (ML) and Neural Network application on clinical data set of Uddanam CKD for prevention and early detection of CKD. As per statistical analysis we can prevent the CKD in the Uddanam area. As per ML analysis Naive Bayes model is the best where the process model is constructed within 0.06 seconds and prediction accuracy is 99.9%. In the analysis of NNs, the 9 neurons hidden layer (HL) Artificial Neural Network (ANN) is very accurate than other all models where it performs 100% of accuracy for predicting CKD and it takes the 0.02 seconds process time.

***Index Terms*: Artificial Neural Network (ANN), Chronic Kidney Disease (CKD), Machine Learning (ML), Statistical Analysis**

I. INTRODUCTION

According to W.H.O., UDDANAM (srikakulam district of A.P., India) is the place with highest concentrated chronic kidney disease among three regions in the world after srilanka and Nicaragua. Uddanam is north coastal region of Andhra Pradesh, India consists of mandals like Kaviti, Sompetha, Itchapiram, Palasa and Vajrapukotturu and there are more than 100 villages in total. It was predicted that more than 4,500 people had died from kidney disease in the last 10 years as per 2015 [16]. Around 34,000 people had kidney disease in uddanam from past 20 years .This disease is mostly affected to farmers and agricultural workers, who are growing coconut and cashew as main crop. Hence automatic diagnosis tools will help the doctor to identify the disease early and quickly and help the patients' survival rate. Classification methods are often utilized in many automatic medical diagnosis tools. Identifying kidney disease at an early stage is a difficult job. By using tool the burden of diagnosis can be reduced [19]. Chronic Kidney Disease (CKD) affects the structure and functionality of kidney. The disease has many stages depending on the level of Glomerular Filtration Rate (GFR) .A prolonged illness may result in complications like anemia, high BP, nerve damage, weak bones, heart, or blood vessel problems, etc. Besra et al., proposed a method to predict CKD with high accuracy and also an estimation of percentage of kidney damage. This helps in diagnosing accurately the stages of CKD and thereby better treatment. It uses classification techniques and Glomerular Filtration Rate (GFR) value calculation. Chronic kidney disease (CKD) is fast growing and associated with extreme threat of cardiovascular and end-stage kidney disease, which are potentially escapable through early discovery and treatment of people in danger. Machine learning algorithm can be used to analyze the disease effectively in the prior stage. Thus, machine-predicted analysis is most popular now days to detect kidney disease. Chronic kidney failure (perpetual kidney disease 'CKD') is a genuine malady that identified with the steady loss of kidney work. It is viewed as one of the wellbeing dangers in the creating and undeveloped nations at beginning times; couple of side effects can be distinguished, where the CKD may not wind up evident until critical kidney work impeded happen. CKD treatment centers on diminishing the kidney harm movement by controlling the fundamental reason, which requires sickness location at introductory stages. In this research work, we used some MLs like Naïve Bayes, Bayes Net, SVM and Logistic algorithms and comparing their performance of Uddanam kidney patient's Data set. As well, we apply Artificial Neural Network (ANN) algorithms also for 100% accuracy of the data set. This work is very useful to area of uddanam people for awareness about kidney disease, though we can prevent and curing the disease easy and earlier. It is also worthy work for predicting the disease in early stage with low cost and less time. This work is use full for doctors and government for Prevalence, Prediction and Prevention of uddanam CKD.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
K. B. Anusha*, Assistant Professor, Department of Computer Science and Engineering ,Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh , India.
Dr. T. PanduRanga Vital (*Corresponding Author), Associate Professor, Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.
K. Sangeeta, Sr. Assistant Professor, Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India.

II. REVIEW OF LETERATURE

In this section, we referred so many research papers from reputed journals like Elsevier; Springer and IEEE transactions related this work. We present some of effective papers as review and authors findings and their model works related this concept.

Chronic Kidney Diseases (CKD) is major disease that it affects the structure and functionality of kidney. The progress of CKD can be prevented by detecting the disease in early stages and specific treatment at different stages. The current techniques for clinical CKD stage diagnoses use the level of GFR mean that glomerular filtration rate. This method often produces inaccurate and contradictory results as there are different approaches to estimate GFR. Fenget et al., (2019) researched on CKD that ML techniques are used to construct high-performance CKD stage diagnosis models to diagnose CKDs stages without estimating GFR. Positive metabolite levels in blood samples were evaluated by MS (mass spectrometry) which constitutes the dataset. Feature Selection method is used to recognize which is nearly correlated metabolite characteristics related to CKD developments. These chose metabolite features was used to construct an improved CKD stage predicting models, improved with respect to diagnosis cost and time when com. Our improved model could achieve over 98% accuracy in CKD prediction. Furthermore, the models and results are validated by applying unsupervised learning algorithms [1]. Internet of things (IOT) and cloud computing assumes a significant role in disease prediction particularly in smart cities. IOT devices generates huge amount of data onto chronic kidney diseases (CKD) that can be stored in the cloud. This massive data can be utilized for incremental performance or accuracy that for prediction of CKD on cloud environment. Abdelaziz et al., (2019) propose a technique for predicting or indicating the CKD on cloud computing environment. It utilizes a hybrid model that assumes two intelligent techniques which are NNs and linear regression (LR). LR is utilized to decide basic factors that impact on CKD. NN is used for prediction purpose. The outcomes demonstrate that, this hybrid intelligent model is 97.8% accurate in predicting CKD. The proposed model is better than the majority of the models referred to in the associated research works by 64% [2]. Chronic Kidney Disease (CKD) affects the structure and functionality of kidney. The disease has many stages depending on the level of Glomerular Filtration Rate (GFR) .A prolonged illness may result in complications like anemia, high bp, nerve damage, weak bones, heart, or blood vessel problems, etc. Besra et al., proposed a method to predict CKD with high accuracy and also an estimation of percentage of kidney damage. This helps in diagnosing accurately the stages of CKD and thereby better treatment. It uses classification techniques and Glomerular Filtration Rate (GFR) value calculation. Chronic kidney disease (CKD) is fast growing and associated with extreme threat of cardiovascular and end-stage kidney disease, which are potentially escapable through early discovery and treatment of people in danger. Machine learning algorithm can be used to analyze the disease effectively in the prior stage. Thus, machine-predicted analysis is most popular now days to detect kidney disease. Hasan et al., discusses an ensemble method based classifier to increase accuracy in classification for kidney disease diagnosis. Ensemble methods uses many learning algorithms to achieve better prediction results compared to any of the learning algorithms used alone. Moreover, tenfold cross-validation is used and performance is analyzed using ROC curve. CKD datasets from the UCI machine learning repository is used for experimental analysis and results achieves the state-of-the-art performance [3].

Hore et al., (2018) worked on CKD, in their work, a genetic algorithm (GA) trained neural network (NN)-based representation model has been proposed to detect CKD which has converted into one of the widely spread diseases across the world. Studies and reviews in various areas of India have proposed that CKD is turning into a noteworthy concern step by step. The money required for the treatment and future results of CKD could be unbearable to many, if not identified at an earlier stage. For this reason, the algorithm NN-GA has been suggested which essentially beats the issue of local search learning based algorithms to train NNs. The input weight vector of the NN is gradually optimized by using GA to train the NN. The model has been contrasted and familiar classifiers like RF Trees, MLP-FFN, and different NNs. Process and valued the various classifiers with accuracy, precision, gradient and F-Measure. The test results recommend that NN-GA model is fit for distinguishing CKD more productively than some other presented model. Right now, health issues progressively encourages the interest of data scientists. In fact, data analytics as a quickly developing zone can be the correct answer for oversee, identify and anticipate illnesses which undermine human life and cause a high monetary expense to health systems [5]. Alaoui et al., (2018) looks to build up a statistical and predictive analysis of an available dataset related to chronic kidney disease (CKD) by utilizing the broadly used software package called IBM SPSS [6][18]. In fact, we figure out how to make a 100%

precise model based on XGBoost linear machine learning algorithm for effective classification of patients into; affected by CKD or not affected. Kriplani et al., were gone through 224 records of CKD accessible on the UCI repository CKD since 2015. Their proposed technique is based on DNN with a precision of 97%. This model shows better results as compared to already existing algorithms as it is using the cross-validation technique to avoid over fitting. This model works efficiently provided the disease is detected at an earlier stage [7]. The table 1 shows the various authors' contributions on CKD. In this we present some of authors' recent research works and findings on classifications and predictions of CKD with related this work. Some of the authors used the UCI data set and some other utilized their local area's CKD data set.

Table 1: Contribution of Authors on Predicting CKD

Ref. No.	Authors Reference	Highlighted Contributions	Area of Application and Description	Year
[7]	Kriplani, et al.,	Chronic kidney disease with an exactness of 97%.	Prediction of Chronic Kidney Diseases Using Deep Artificial Neural Network Technique	2019
[1]	Abdelaziz et al.,	NN-LR hybrid intelligent model (97.8%) accurate	Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities	2019
[4]	Hasan et al.,	Ensemble method based classifier (91-97%)	Prediction of Chronic Kidney Disease	2019
[13]	Chaitanya et al.,	KNN (99.9%)	Detection of Chronic Kidney Disease by Using Artificial Neural Networks and Gravitational Search Algorithm	2019
[6]	Alaoui et al.,	IBM SPSS (100%) and 100% precise model based on XGBoost	Statistical and Predictive Analytics of Chronic Kidney Disease	2018
[12]	Ravindra et al.,	Accuracy SVM (94.44%)	Classification of non-chronic and chronic kidney disease using SVM neural networks	2018
[11]	Boukenze et al.,	SVM (91-95%)	Predicting Chronic Kidney Failure Disease Using Data Mining Techniques	2016
[8]	Chatterjee et al.,	NN-MCS (91-95%)	Hybrid modified Cuckoo Search-Neural Network in chronic kidney disease	2017

Chronic kidney failure (perpetual kidney sickness 'CKD') is a genuine malady that identified with the steady loss of kidney work. It is viewed as one of the wellbeing dangers in the creating and undeveloped nations at beginning times; couple of side effects can be distinguished, where the CKD may not wind up evident until critical kidney work impeded happen [21]. CKD treatment centres on diminishing the kidney harm movement by controlling the fundamental reason, which requires sickness location at introductory stages. Chatterjee et al., proposed classifiers execution is estimated regarding distinctive execution measurements. The test results portrayed that the NN-MCS can distinguish CKD all the more productively contrasted with some other existing model. Chronic kidney disease (CKD) is a noteworthy general wellbeing worry with rising predominance. Salekin et al., examinations, they think about 24 prescient parameters and make an AI classifier to distinguish CKD. Salekin et al., assess their methodology on a dataset of 400 people, where 250 of them have CKD. Utilizing their methodology we accomplish a recognition precision of 0.993 as per the F1-measure with 0.1084 root mean square blunder. This is a 56% decrease of mean square mistake contrasted with the cutting edge (i.e., the CKD-EPI condition: a glomerular filtration rate estimator). We likewise perform highlight choice to decide the most applicable traits for recognizing CKD and rank them as per their consistency. We recognize new prescient qualities which have not been utilized by any past GFR estimator conditions. At last, we play out a cost-precision trade-off investigation to distinguish another CKD recognition approach with high exactness and ease [8].

CKD is a worldwide general medical issue, influencing around 10% of the populace around the world. However, there is minimal direct proof on how CKD can be analyzed in a methodical and programmed way [20][21]. Subasi et al., (2017) researches how CKD can be analyzed by utilizing AI (ML) strategies. ML calculations have been a main thrust in identification of irregularities in various physiological information, and are, with an incredible achievement, utilized in various characterization assignments. In the present investigation, various diverse ML classifiers are tentatively approved to a genuine informational collection, taken from the UCI Machine Learning Repository, and their discoveries are contrasted and the discoveries revealed in the ongoing writing. The outcomes are quantitatively and subjectively examined and their discoveries uncover that the irregular woods (RF) classifier accomplishes the close ideal exhibitions on the recognizable proof of CKD subjects. Thus, we demonstrate that ML calculations serve significant capacity in conclusion of CKD, with acceptable heartiness, and their discoveries propose that RF can likewise be used for the finding of comparative diseases [10]. Kidney failure malady is being seen as a genuine test to the therapeutic field with its effect on a gigantic populace of the world. Without side effects, kidney illnesses are regularly recognized past the point of no return when dialysis is required direly [12][16]. Boukenze et al., (2016) examined that propelled Data mining advancements can help give choices to deal with this circumstance by finding concealed examples and connections in medicinal information. The target of this exploration work is to foresee kidney ailment by utilizing numerous AI calculations that are SVM, MLP, C4.5, BN and K-NN. The point of this work is to think about those calculations and characterize the most productive one(s) based on numerous criteria. The database utilized is "Perpetual Kidney Disease" actualized on the WEKA stage. From the test results, it is seen that C4.5 and MLP have the accurate precision values. Be that as it, when contrasted and Receiver Operating Characteristic (ROC) curve, C4.5 gives off an impression of being the most productive [11].

III. PROPOSAL MODEL

The fig.1 shows the Kidney Disease (KD) predicting proposal model. In this process, firstly we collected total 1055 records from the clinical centers of different hospitals and people of UDDANAM area in Srikakulam District of Andhra Pradesh India from 2016 to 2019.

In this, 655 instances of CKD and 399 instances of non-CKD clinical information is recorded with good questionnaire related to Kidney Diseases. Particularly, we focused on 37 feature attributes of CKD and non-CKD that are Age, Sex, Blood pressure, specific gravity, Albumin, Sugar, mcv, Platelet count, Pus cell clumps, Blood Glucose, Blood urea, Neutrophils, Red blood cells, Bacteria, potassium, Chloride, Hemoglobin, packed cell, White blood cells, Pus cell, Lymphocytes, Eosinophils, Monocytes, Basophiles, Bilirubin, Red blood cells count, Anemia, Coronary Artery, , serum, urea, Creatinine, Uric Acid, Diabetes, Hypertension and so on. We classify the data set with a class CKD refers to 1 and non-CKD refers to 0.

The whole information is put into the data table and converted into *.CSV format for preprocessing. In this, we apply different Data Mining preprocessing techniques. The preprocessed data is utilized for statistical analysis and performance analysis of the chronicle kidney disease (CKD) with utilizing MLs and NNs. As per analysis we choose the best algorithm for detecting CKD in early and accurate. These reports are very important and useful for the doctors and analysts for identify the CKD with less time and low cost. This system is automated and gives the reports automatically with in fraction of minutes.

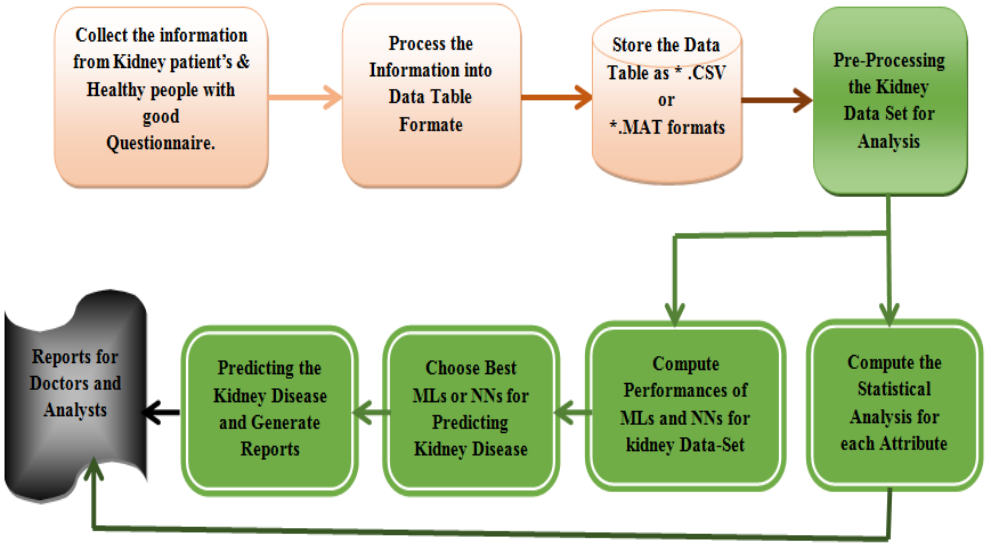

Fig. 1: Prediction Chronic Kidney Disease Model

Naive Bayes Classification:

Naive Bayes is a classifier, classifies with utilizing probability. It can utilize the kernel estimator but not meet for discretization. The conditional probabilities are base of this process[15]. Naive Bayes utilizes the Bayes theorem and it was formulated with probability. Bayes theorem determines the probability shown in equations.

$$P(S/X)=\frac{P(X/S)P(S)}{P(S)} \quad \text{.....(1)}$$

$$P(C/X)=P(X_1|C)\times P(X_2|C)\times.....\times P(X_n|C)\times P(C) \quad \text{.....(2)}$$

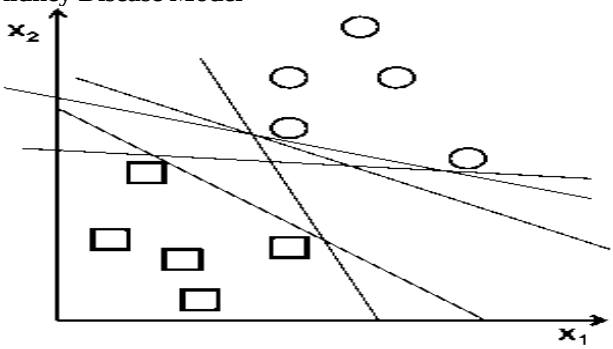

Fig. 2: SVM with Classes

Support Vector Machine (SVM):

SVM is a associated supervised ML model that classifies the data set utilized for regression analysis and classification. A SVM is discriminatory classifier formally determined by a dividing hyper plane. In other, the given marked training data in a hyper plane the algorithm produces optimized hyper plane which classifies new models. A SVM model is the portrayals of the instances of the various classifications are separated by comprehensible space that is as an extensive as possible [17]. New instances or precedents are then mapped or represented into that equivalent gap and predicted to have a place with a classification dependent on which side of the hole they fall. As per Linearly separable set of 2D data points which have a place toward one of two classes, locate a separating straight line.

We can characterize a model to evaluate the value of the lines:

A line is awful in the event that it passes excessively near the point since it won't generalize effectively. In this way, our objectives ought to be to discover the line going beyond what many would consider possible from all points. At that point the activity of SVM depends on finding the hyper plane that gives the biggest least separation to the preparation models. Twice, this separation gets the significant name of edge inside SVM's hypothesis. Subsequently, the optimal separating hyper plane boosts training information [24].

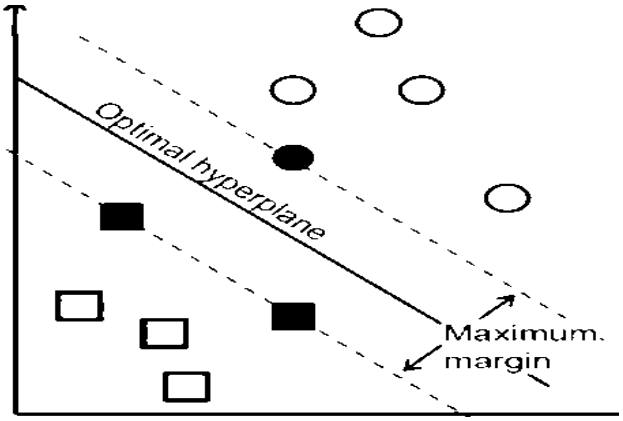

Fig. 3: SVM with optimal hyper plane

Hyper plane (HP) equation is shown in below:

$$f(x) = \beta_0 + \beta^T(x) \text{.....(3)}$$

$\beta 0$ Represents the Bias and β represents the weighted vector of the problem. The best HP can be presented in different ways in infinitely by extent of β and β_0. Among all the feasible internal representation of the HP is one. The equation is

$$|\beta_0| + |\beta^T x| = 1 \text{.....(4)}$$

Where, x is training data that is nearby by the HP. In usual the training points that are closest to the HP are called support vectors. This illustration is recognized as the canonical HP. Now we can determine the distance as below equation.

$$Distance = \frac{|\beta_0 + \beta_x|}{\|\beta\|} \text{.....(5)}$$

For the canonical HP, the numerator value is equivalent to one. Then, the support vector distance is measured as

$$Distance_{\text{sup portvectors}} = \frac{|\beta_0 + \beta_x|}{\|\beta\|} = \frac{1}{\|\beta\|} \text{.....(6)}$$

So, we can measured M is equal to twice the distance to the closest instances

$$M = \frac{2}{\|\beta\|} \text{.....(7)}$$

Confusion matrix:

Data set S = {e_1, e_2… e_i … e_n} contains 'n' data elements with two class of groups. So, Table1 shows the confusion matrix with rows and columns (two by two matrixes used in cancer prediction study) that are p, q, r and s. The 'p' indicates true-positive (TP) classified data elements or instances of the total 'n' elements data set. In addition, 'q' indicates FP (false-negative), 'r' specifies FP (false-positive) and's' represents TN (true-negative) data points or examples of data set.

Table 2: Confusion matrix

Actual Result	Detected result	
	TruePositive(TP) P	FalseNegative(FN) q
	FalsePositive(FP) R	TrueNegative(TN) s

Accuracy is measured with using true values that are p (TP) and s (TN). Accuracy is estimated by the Equation (8).

$$Accuracy = \frac{p+s}{p+s+q+r} \text{.....(8)}$$

Precision is described as the fraction of a cluster that consists of elements of specified class.Precision is ratio of the true positives among the cluster by addition of TP (p) and FP (r). Precision is calculated by the Equation (9)

$$Precision = \frac{p}{p+r} \text{.....(9)}$$

Recall is defined as probability of related objects or elements (data points) which is selected from the specified class. In other words, recall is described as combinations of all elements that are grouped in to a specific class. Recall is a procedure of the proper classification of data (TPs) and misclassified data (FPs). So, recall is computed by Equation (10)

$$Recall = \frac{p}{p+q} \text{.....(10)}$$

Receiver Operator Characteristic Curve:

The ROC curve is used in graphical representation between TP Rate (TPR) and FP Rate (FPR) of a given dataset. In the ROC curve axis X represent the FPR, it shows the clusters which has negative incorrect group or total negative group in given dataset. The axis Y represents TPR and it shows the total positive group in given dataset. A ROC graph is utilized to classify and project the performance and accuracy of an algorithm. It is usually utilized as a part of machine learning, medical decision making and DM research group (Fawcett 2003). The ROC curve represents the performance of algorithm with respect to FPR and TPR respectively for all the datasets. The ROC curve resides (0, 0) means each object of the cluster declared in negative class. If curve resides (1, 1) then each object of the cluster declared in positive class and if it resides (1, 0) means clusters are ideal position.

Artificial Neural Network:

Fig.4 shows the ANN feed-forward network system. Consider a sample for the two-layered network. Bring an output layer with two units start to finish, a hidden layer with four units and system with 'n' input units [14][23]. The 'n' inputs appear as circles and they do not fit in with any layer of the system. Any layer that is not an output layer is a hidden layer. This system consequently has 1 hidden layer and one destination layer. Here it demonstrates every one of the associations between the units in various layers. Here it indicates how every layer is exclusively associated with the previous layer.

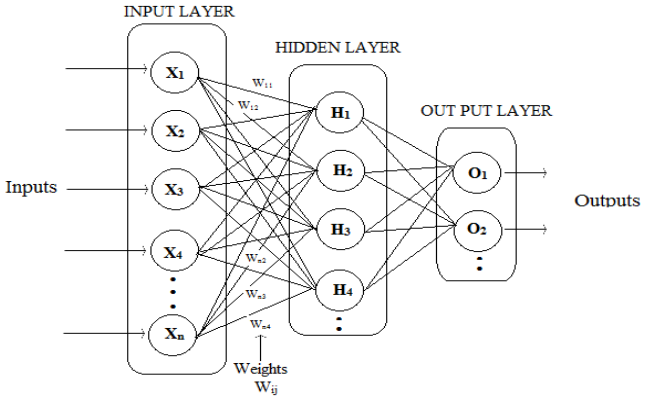

Fig. 4: ANN Feed Forward system

IV. RESULTS AND DISCUSIONS

Statistical Analysis on Uddanam CKD data set:

As per statistical analysis, we refer to 37 attributes(Age, Sex, Blood pressure, specific gravity, Albumin, Sugar, mcv, Platelet count, Pus cell clumps, Blood Glucose, Blood urea, Neutrophils, Red blood cells, Bacteria, potassium, Chloride, Hemoglobin, packed cell, White blood cells, Pus cell, Lymphocytes, Eosinophils, Monocytes, Basophiles, Bilirubin, Red blood cells count, Anaemia, Coronary Artery, , serum, urea, Creatinine, Uric Acid, Diabetes, Hypertension and so on) . In this, the main attributes related to CKD are focussed and gives the analysis about CKD and NON-CKD values. Most of the attribute statistics are measured with Min, Max, Mean and Median Values. Some of the attribute statistic parameters are present or not present, Normal or Abnormal, Male or female and Yes or NO values. The table describes the CKD and NON-CKD differences in detail. We notify some of the results that the Blood Pressure is very high and some have low in CKD patients (high B.P -41.5% and low B.P. -41.3%) rather Non- CKD patients. The BP is related to cause of chronic kidney disease. The attribute Albumin values of CKD is very high values (Mean 4.57±10.87) than non-CKD values (Mean 0.18±0.54). The blood glucose and blood urea are also causes of CKD that it is more difference than non-CKD. CKD is impacted by the red blood cells report that abnormal RBCs are heavy cases in CKD (43.8%) than non-CKD. As well Bacteria present 33.8% in CKD cases and only 3% present in non-CKD.

Table 3: Statistical Analysis of CKD and NON-CKD for UDDANAM Data Set

Attributes	Category	Statistical values of CKD	Statistical values of NON_CKD	Attributes	Category	Statistical values of CKD	Statistical values of NON_CKD
Age	Min Mean Median Max	2.00 52.15±16.08 55.00 90.00	4.00 46.94±15.95 47.00 80.00	**Neutrophils**	Min Mean Median Max	1.40 59.17±13.15 61.20 84.90	1.40 59.42±13.30 61.40 84.90
Sex	Male Female	490(74.7%) 166(25.3%)	301(75.4%) 98(24.6%)	**Red blood cells**	Normal abnormal	369(56.2%) 287(43.8%)	370(92.7%) 29(7.3%)
Blood pressure	Normal High low	113(17.2%) 272(41.5%) 271(41.3%)	260(65.2%) 8(2.0%) 131(32.8%)	**Bacteria**	Present NotPresent	222(33.8%) 434(66.2%)	12(3.0%) 387(97.0%)
specific gravity	Min Mean Median Max	1.00 1.01±0.00 1.01 1.02	1.00 1.02±0.00 1.02 1.02	**potassium**	Min Mean Median Max	3.00 6.03±8.35 4.90 116.30	3.30 4.37±0.59 4.50 5.50
Albumin	Min Mean Median Max	1.44 4.57±10.87 3.47 14.00	0.00 0.18±0.54 0.00 5.00	**Chloride**	Min Mean Median Max	10.20 106.83±8.82 105.80 144.70	93.60 103.31±4.81 102.30 121.00
Sugar	Min Mean Median Max	0.00 1.57±1.777 1.00 5.00	0.00 0.09±0.29 0.00 1.00	**Hemoglobin**	Min Mean Median Max	4.00 9.37±2.88 8.70 26.20	10.30 14.98±1.53 15.00 18.50
mcv	Min Mean Median Max	7.40 81.21±10.63 82.00 110.00	49.00 80.84±8.55 80.00 100.00	**packed cell**	Min Mean Median Max	10.70 26.73±7.40 25.00 50.30	29.00 45.31+4.89 45.00 54.00
Platelet count	Min Mean Median Max	0.10 2.31±0.92 2.20 7.90	1.00 2.42±0.84 2.30 7.90	**White blood cells**	Min Mean Median Max	1290.0 7050.05±375.8 6560.00 58802.00	3200.00 7551.88±1908.55 7300.00 13200.00
Pus cell clumps	Present Not present	435(66.3%) 221(33.7%)	20(5.0%) 379(95.0%)	**Pus cell**	Normal abnormal	334(50.9%) 322(49.1%)	371(93.0%) 24(7.0%)
Blood Glucose	Min Mean Median Max	22.00 180.29±88.21 156.00 490.00	70.00 114.72±33.16 111.00 312.00	**Lymphocytes**	Min Mean Median Max	10.50 30.79±10.83 29.20 68.90	10.50 30.75±10.93 29.20 68.90

Blood urea	Min Mean Median Max	1.50 72.30±57.87 53.00 391.00	10.00 35.37±14.94 34.00 95.00	**Eosinophils**	Min Mean Median Max	0.00 3.78±3.97 2.60 22.50	0.00 3.64±3.80 2.60 22.50
serum	Min Mean Median Max	0.50 3.29±4.34 2.40 48.10	0.40 1.01±0.60 0.90 5.90	**Monocytes**	Min Mean Median Max	0.60 5.55±3.39 4.80 44.00	0.60 5.45±3.22 4.80 44.00
urea	Min Mean Median Max	10.96 204.74±1812.12 77.94 27066	10.96 217.52±1906.06 77.94 27066.00	**Basophils**	Min Mean Median Max	0.00 0.13±0.20 0.10 1.80	0.00 0.13±0.20 0.10 1.80
Creatinine	Min Mean Median Max	0.26 26.09±335.30 7.93 6089.00	0.26 22.78±304.09 7.83 6089.00	**Bilirubin**	Min Mean Median Max	0.02 0.73±1.26 0.49 18.22	0.02 0.71±1.21 0.48 18.22
Uric_Acid	Min Mean Median Max	0.22 6.73±2.08 6.76 16.54	0.22 6.68±2.11 6.75 16.54	**Red blood cells count**	Min Mean Median Max	1.27 3.42±0.89 3.29 6.29	2.10 5.19±0.77 5.20 6.50
Diabetes	Yes no	396(60.4%) 260(39.6%)	9(2.3%) 390(97.7%)	**Anemia**	Yes no	302(46.0%) 354(54.0%)	13(3.3%) 386(96.7%)
Hypertension	Yes no	396(60.4%) 260(39.6%)	15(3.8%) 384(96.2%)	**Coronary Artery**	Yes no	191(29.1%) 465(70.9%)	17(4.3%) 382(95.7%)

Classification Algorithms Comparison:

The table 4 shows the accuracy comparision analysis for different algorithms. In this, we mesure the accuracy of the different ML algorithms with the accuracy parameters like FP rate, TP rate, ROC and precision values. The Naïve Bayes is the best algorithm rather oter algorithms where the ROC value is 0.999 and built the model within 0.06 seconds. The next best algorithms is Logistic algorithm (The ROC value is 0.996). The second best in built the model is Bayes Net where it takes 0.14 seconds. The choosing all four algorithms are very accurate for the CKD uddanam data set where all ML models ROC values are above 0.976 and the maximum of the built-in models time is below 0.28 seconds.

Table 4: Comparison Accuracy Analysis for all MLs with using Accuracy Measures like TP, FP, Precision and ROC Values

Classifier	Built the Model in sec	Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Naive Bayes	0.06	CKD	0.997	0.038	0.978	0.997	0.987	0.999
		NCKD	0.962	0.003	0.995	0.962	0.978	0.999
		AVG	**0.984**	**0.025**	**0.984**	**0.984**	**0.984**	**0.999**
SVM	0.28	CKD	0.985	0.033	0.980	0.985	0.983	0.976
		NCKD	0.967	0.015	0.975	0.967	0.971	0.976
		AVG	**0.976**	**0.024**	**0.977**	**0.976**	**0.977**	**0.976**
Logistic	0.26	CKD	0.991	0.023	0.986	0.991	0.989	0.995
		NCKD	0.977	0.009	0.985	0.977	0.981	0.998
		AVG	**0.986**	**0.017**	**0.986**	**0.986**	**0.986**	**0.996**
BayesNet	0.14	CKD	0.992	0.023	0.986	0.992	0.989	0.989
		NCKD	0.977	0.008	0.987	0.977	0.982	0.989
		AVG	**0.987**	**0.017**	**0.987**	**0.987**	**0.987**	**0.989**

The table 5 shows the comparison analysis of the all used ML models. The confusion matrix is constructed with two classes that are CKD and N-CKD. The total data set elements are 1055. In this the total CKD elements are 656 and N-CKD data elements are 399. As per analysis of Naïve Bayes, the class CKD 654 data elements are classified as true positive(TP) of the total 656 CKD elements and remaining 2 elements are classified incorrectly that in true negative (TN). In addition, 384 elements are classified as FP (false-positive) and 15 data points specifies FN (false-negative). As well remaining MLs also constructed the confusion matrix for each their TPositive, TNegative, FPositive and FNegative values.

Table 5: Comparison Accuracy Analysis for all MLs with using Confusion Matrix

ML-Algorithm	Confusion Matrix	Class			ML-Algorithm	Confusion Matrix	Class		
Naive Bayes	Class	CKD	N-CKD	Total	Logistic	Class	CKD	N-CKD	Total
	CKD	654	2	656		CKD	650	6	656
	N-CKD	15	384	399		N-CKD	9	390	399
	Total	669	386	1055		Total	659	396	1055
SVM	CKD	646	10	656	Bayes Net	CKD	651	5	656
	N-CKD	13	386	399		N-CKD	9	390	399
	Total	659	396	1055		Total	660	395	1055

The fig.5 shows the comparative analysis of MLs with utilizing the ROC values. As per analysis the least perform ML is SVM with 0.976 accuracy and the best perform ML model is the Naïve Bayes with ROC value 0.999.

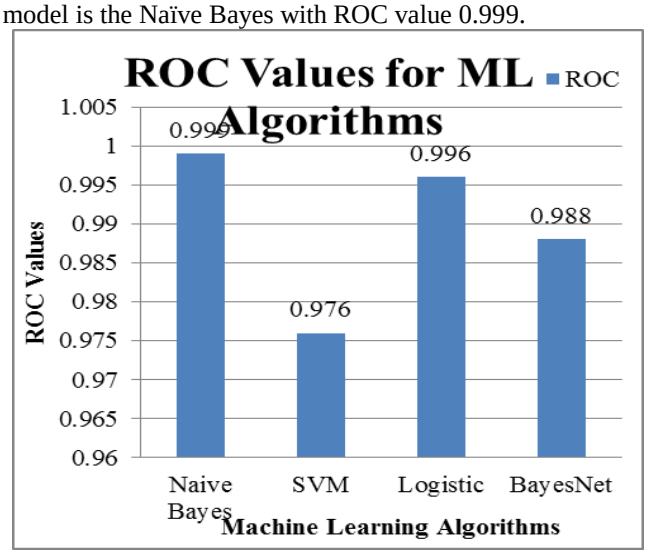

Fig. 5: Comparative ML Models Analysis with ROC values

The fig.6 shows the comparative analysis of MLs with utilizing the time complexity that built the model in second. As per analysis, takes the least time for built ML model is Naïve Bayes with 0.06 seconds and maximum time taking built ML is SVM with 0.28 seconds. As per comparative analysis the best algorithms is Naïve Bayes to other ML models.

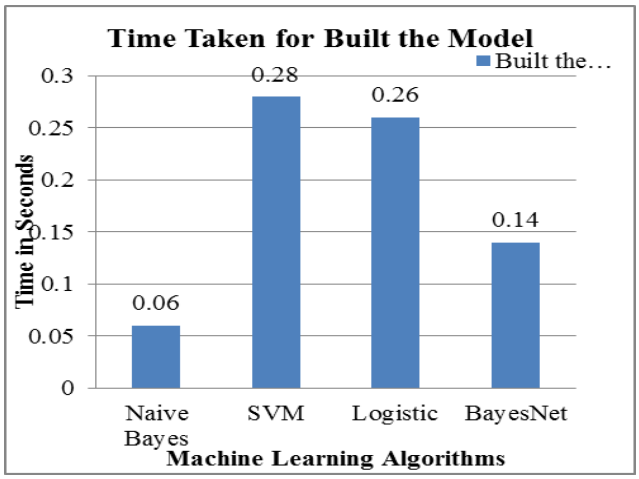

Fig. 6: Time Complexity of ML Models

Artifitial Neural Network (ANN) Algorithm Analysis on Uddanam CKD Data Set:

The total uddanam CKD data is divided into Training and Target data sets. The Training data set contin 37 feature attributes. The Target data is one dimentional data contain 1 for CKD and 0 for non-CKD. The ANN is constructed with three (Input, Hidden and Output) layers. The input layer contains the 37 feature neurons , the incremental decition nodes or Hidden neurons(7 to 9) are allocated to the Hidden layer (HL). In this experiment we increment the HL neurons 7 to 9 and anlyze the performance of the data set in each stage. In this analysis the total data set instances are divided into two parts randomly that are Training and Testing. The total data set instances are 1055. The training part is 80% (844 instances) and the Testing part is 20% (211) of the total data set.

ANN with 7 Hidden Neurons Classification Analysis:

The fig.7 shows the Best Training performance of the 7 HL Neurons of ANN analysis. The X-axis specifies the epoche number and the Y-axis specifies the Mean Squared Error (MSE). The best training performacne of the data set is 7.854e-10 at epoch 184. The blue colored line shows the Training performance and the Red colored line depicts the Testing-performance.

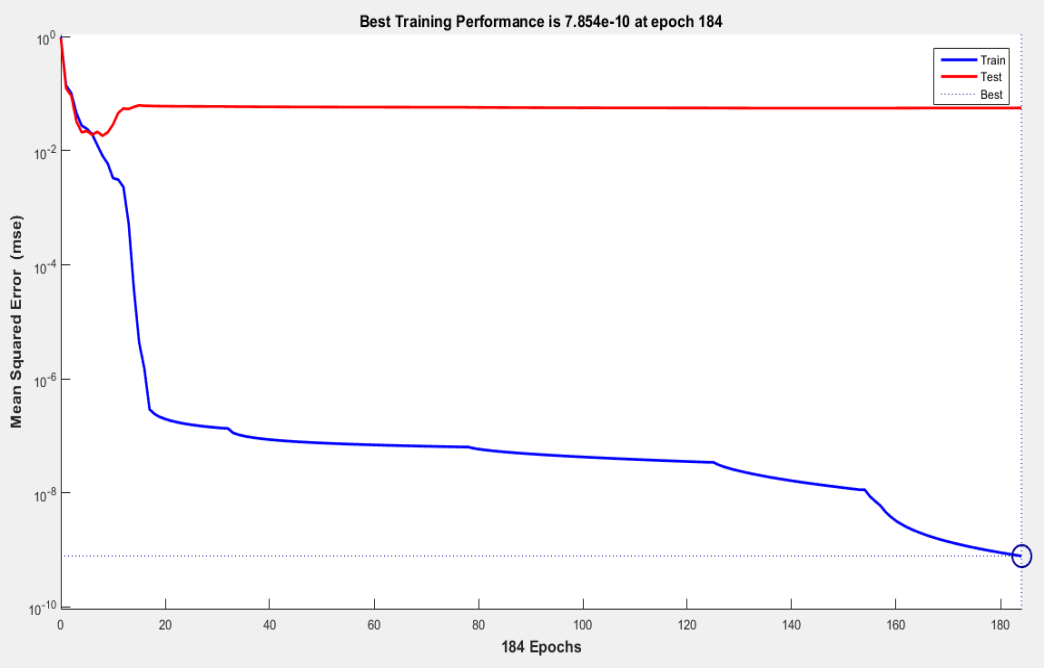

Fig.7: Best Training and Testing Performance Analysis for 7 Neurons HL ANN

The fig.8 shows the statistical values of regression (Train, Test and Total) at the stage of 7 HL neurons of ANN. In this the X-axis indicates the Target values (0 to1) and the output value between 0 and 1. The diagonal line indicates the Fitness of the data points with utilizing linear regression algorithm. The training data set R value is 1 and the testing data set R value is 0.88126 and the total data set R value is 0.98801.

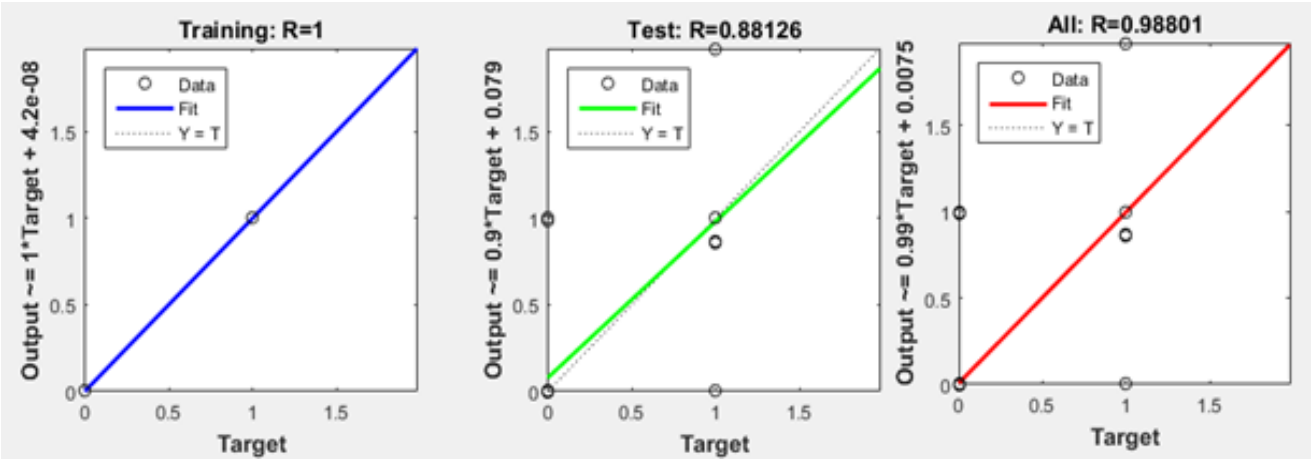

Fig. 8: ANN HL 7 Neurons Linear-Regression (R) values of Train, Test and Total CKD data set

The fig.9 depicts the Gradient and Mu values for 184 iterations. The X-axis shows the epochs and the Y-axis show the gradient value and Mu value. The gradient value is 9.5393e-08 and the Mu value is 1e-10 at epoch 184 for ANN HL 7 neurons analysis. The Fig.10 shows the Error-Histogram of Uddanam CKD Data-set at stage of 7HL neurons ANN. The X- axis indicates the error values that the error values are calculated as Target values minus Output values. The blue color specifies the Training instances error values and the red color indicates the Testing instances of the data set. The Orange color line indicates the Zero Error. Some of less instances are large error rate that value is -0.95 and some less instances error values is +0.15. Most of the instances are nearer to zero values that are error values are -0.05 and +0.05001.

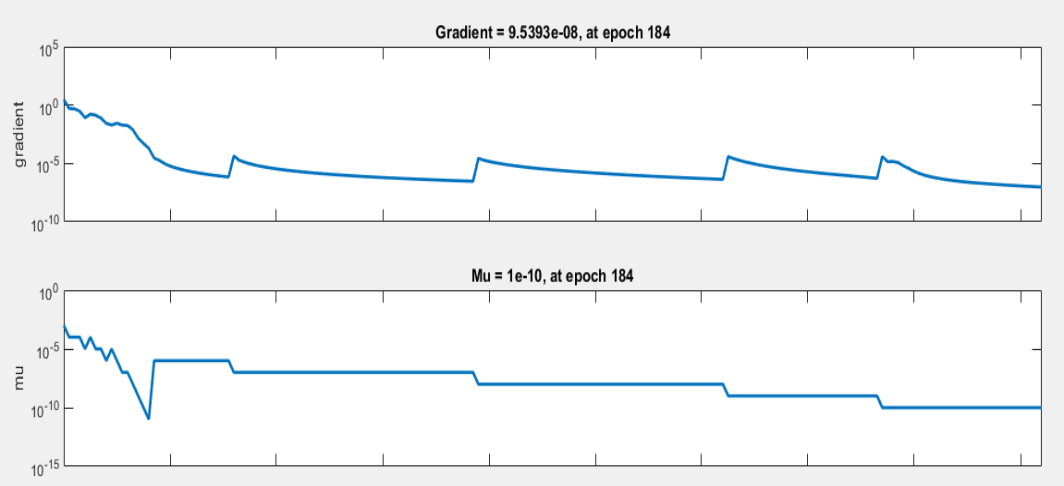

Fig. 9: ANN HL 7 neurons Gradient and Mu Values of Uddanam CKD data set

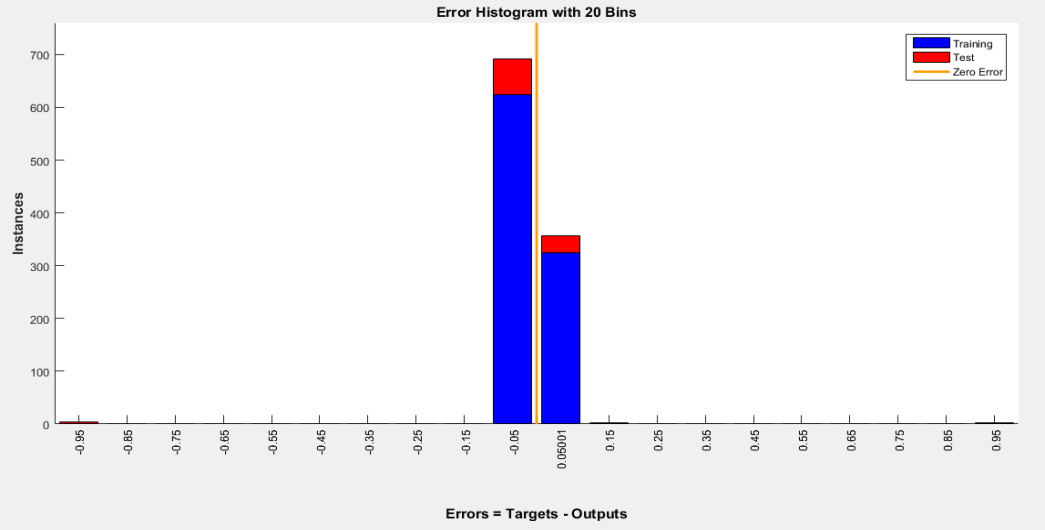

Fig 10: ANN HL 7 Neurons Error-Histogram of Uddanam CKD Data Set

ANN with 8 Hidden Neurons HL Classification Analysis: The fig.11 shows the Best Training performance of the 8 HL Neurons of ANN analysis. The X-axis specifies the epoch number and the Y-axis specifies the Mean Squared Error (MSE).

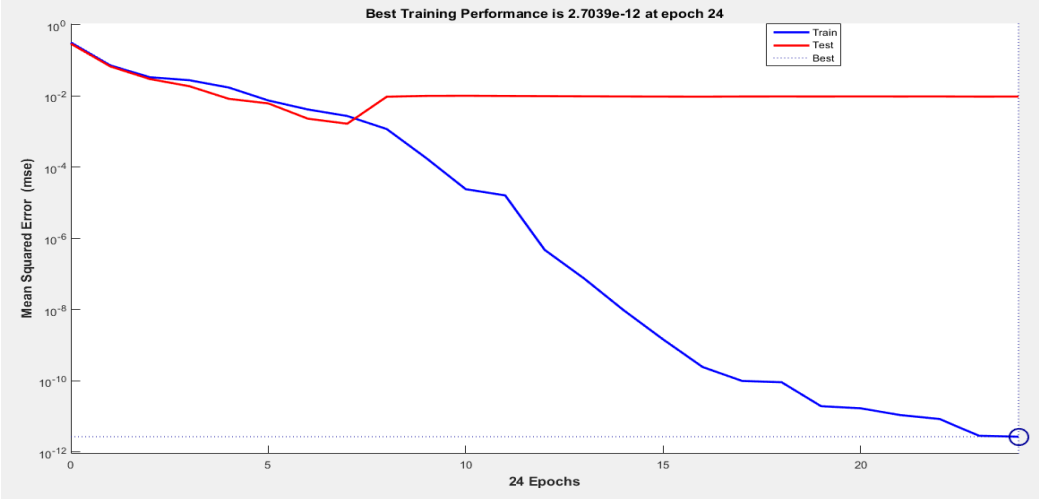

Fig. 11: Best Training and Testing Performance Analysis for 8 Neurons HL

The best training performance of the data set is 2.7039e-12 at epoch 24. The blue colored line shows the Training performance and the Red colored line depicts the Testing performance. The fig.12 shows the statistical values of regression (Train, Test and Total) at the stage of 8 HL neurons of ANN.

In this the X-axis indicates the Target values (0 or 1) and the output value between 0 and 1. The diagonal line indicates the Fitness of the data points with utilizing linear regression algorithm. The training data set R value is 1, the testing data set R value is 0.99807 and the total data set R value is 0.9998.

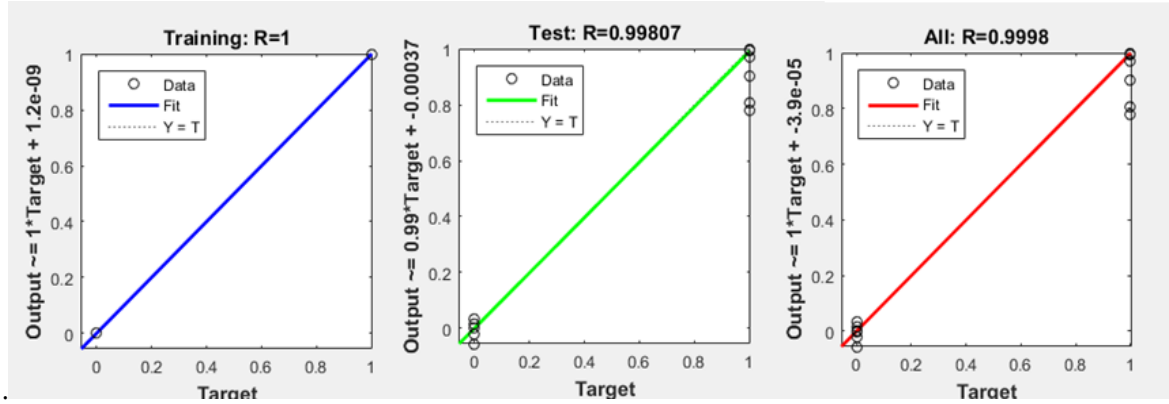

Fig.12: ANN HL 8 Neurons Linear-Regression (R) values of Train, Test and Total CKD Data Set

The fig.13 depicts the Gradient and Mu values for 24 iterations. The X-axis shows the epochs and the Y-axis show the gradient value and Mu values. The gradient value is 7.2291e-08 and the Mu value is 1e-14 at epoch 24 for ANN HL 8 neurons analysis.

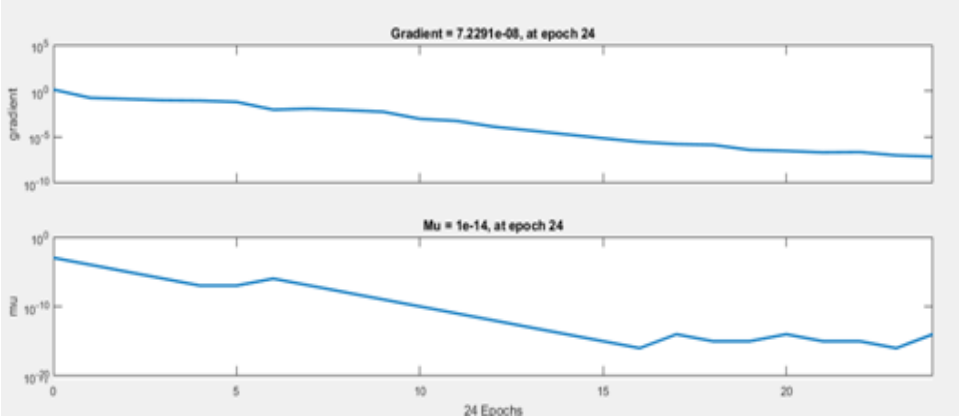

Fig. 13: ANN HL 8 Neurons Gradient and Mu Values of Uddanam CKD data set

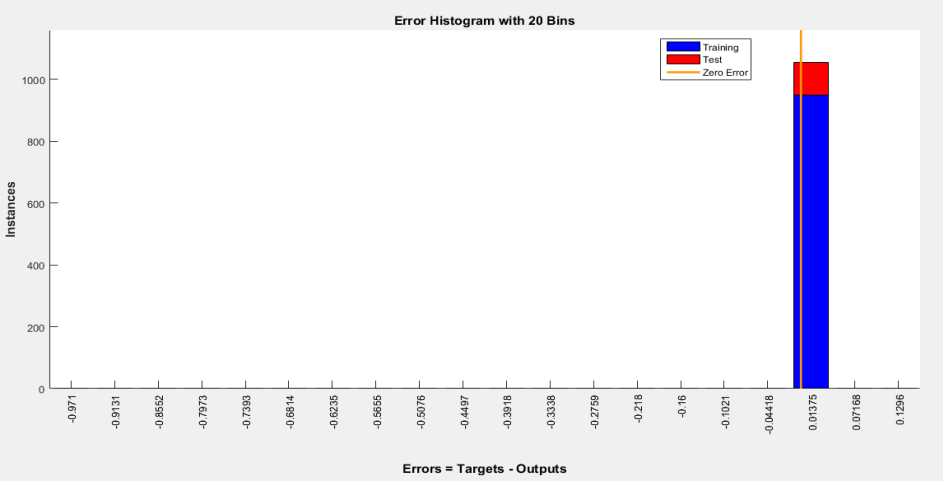

Fig. 14: ANN HL 7 Neurons Error-Histogram of Uddanam CKD Data Set

The Fig.14 shows the Error-Histogram of Uddanam CKD Data-set at stage of 8 HL neurons ANN. The blue color specifies the Training instances error values and the red color indicates the Testing instances of the data set. The orange color line indicates the Zero Error. In this, all the instances are nearer to zero error value that error values are nearer to +0.01375.

ANN with 9 Hidden Neurons Classification Analysis:

The fig.15 shows the Best Training performance of the 9 HL Neurons of ANN analysis. The X-axis specifies the epoch number and the Y-axis specifies the Mean Squared Error (MSE).

The best training performance of the data set is 2.5748e-13 at epoch 20. The blue colored line shows the Training performance and the Red colored line depicts the Testing performance.

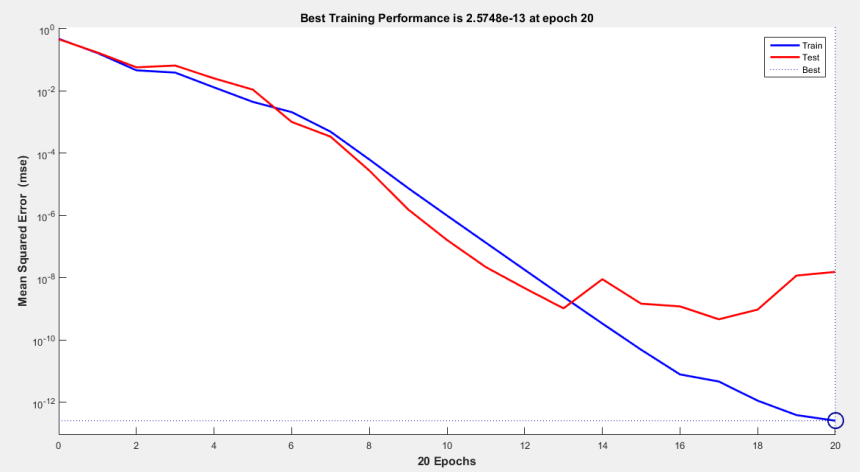

Fig. 15: Best Training and Testing Performance Analysis for 9 HL ANN on CKD Data Set

The fig.16 shows the statistical values of regression (Train, Test and Total) at the stage of 9 HL neurons of ANN. The training data set R value is 1 and the testing data set R value is 1 and the total data set R value is 1. So ANN 9 HL model very accurate for prediction of CKD.

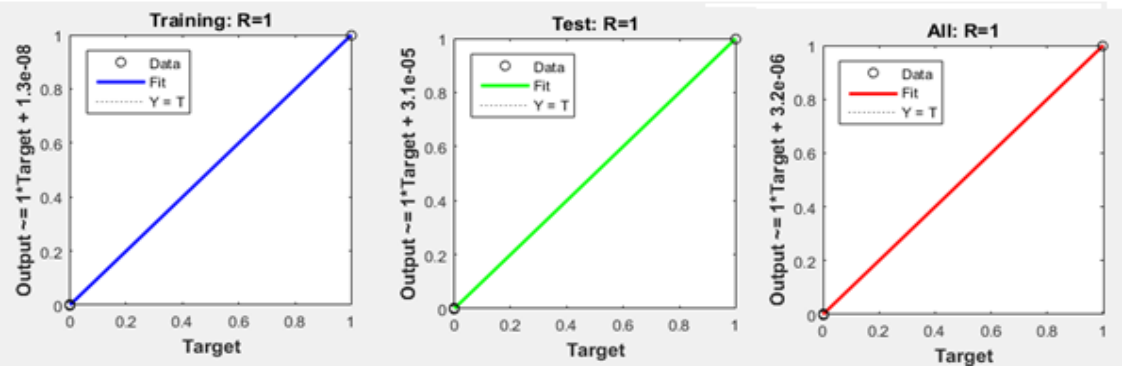

Fig. 16: ANN HL 9 Neurons Linear-Regression (R) values of Train, Test and Total CKD Data Set

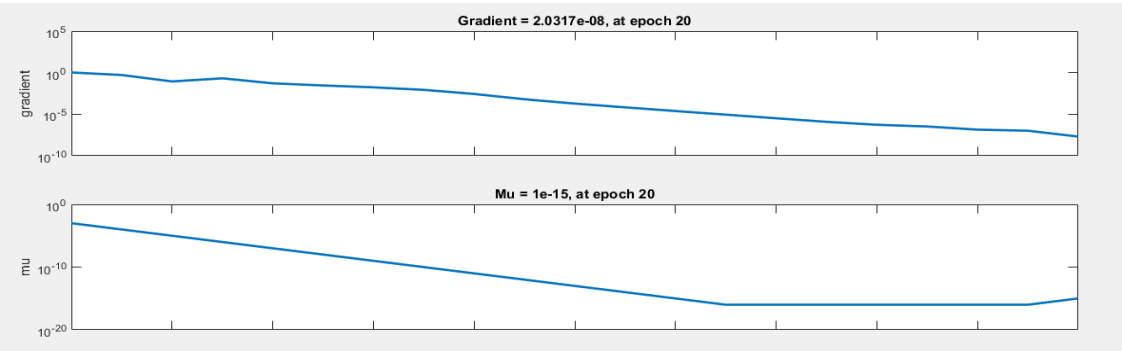

Fig. 16: ANN HL 7 neurons Gradient and Mu Values of Uddanam CKD Data Set

The Fig.17 shows the Error-Histogram of Uddanam CKD Data-set at stage of 9HL neurons ANN. The blue color specifies the Training instances error values and the red color indicates the Testing instances of the data set. The orange color line indicates the Zero Error. In this, all the instances are nearer to zero error value that error values are nearer to -2.7e-05.

Comparative Analysis for 7 to 9 HL Neurons of the ANN:

As per analysis the best performance model is 9 HL ANN where the regression value is 1 and time complexity is very less that it completes the process within 0.02 seconds. The MSE value is very less than 7 and 8 neurons of ANN. So, we conclude the 9 neurons HL of ANN model is the best for prediction UDDANAM CKD with less time and low cost. The table shows the detailed comparative analysis of seven to nine HL of ANN.

Table 6: Comparison Accuracy Analysis for seven to Nine HL of ANN

ANN HL Neurons	Epoch or Iterations	Time Taken for Model	Best performance (MSE Value)	Gradient Values	Mu Values	Regression (R) Value for Total
Seven	184	0.07	7.85e-10	9.54e-08	1.00e-10	0.98801
Eight	24	0.04	2.70e-12	7.23e-14	1.00e-14	0.9998
Nine	20	0.02	2.57e-13	2.03e-08	1.00e-15	**1**

Fig.17 shows the process time for PNN with 7 to 9 HL neurons. In this analysis the 9 neurons HL PNN process time is 0.02 seconds in least comparative others.

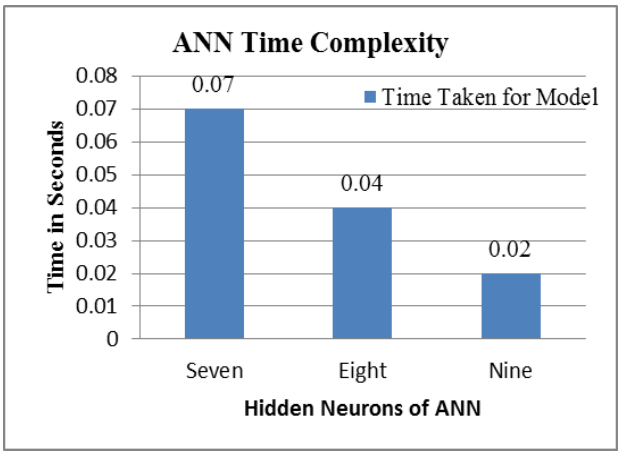

Fig. 17: Time Complexity of 7 to 9 HL ANN Model

The fig.18 shows the one of the accuracy parameters Regression (R) values for each 7 to 9 HL neurons of ANN. The highest performance model is 9 neurons HL ANN that the R value is 1. As per experimental testing results, it is very accurate (100%) model that all the test case values equal to the target values.

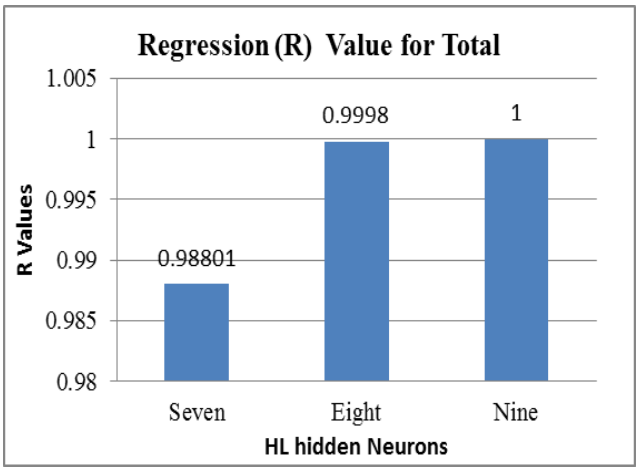

Fig. 18: Time Regression® Value of 7 to 9 HL ANN Model

V.CONCLUSION

The collected Uddanam CKD data set is classified with ML algorithms accurately. The used all ML models (Naive Bayes, SVM, Logistic and Bayes Net) are very useful to predict the CKD where they perform above 98% accuracy. As per comparative analysis the Naïve Bayes the best model in MLs where the ROC value is 0.999 for the prediction. The good statistical results give the highly clinical impact factors of occurring CKD. It is very useful to doctors and averring the people about CKD. In the conclusion, the ANN of 9 neurons HL is the best method for predicting CKD in early stage with low cost where the performance of this model is 100% (Regression (R) values is 1). We will enhance this work linked with environmental and habitual factors of Uddanam occurring of the CKD for this area.

VI. ACKNOWLEDGMENT

We would like to thanks the Director Prof. V.V. Nageswara Rao, Principal Dr. A. Satya Srinivasa Rao and R&D director Dr. K.B.Madhu Sahu and management of AITAM College for give the assistance and support for this work. We are very much thankful to CKD and non-CKD patients' data set belongs to Uddanam area, srikakulam, Andhra Pradesh, India for accepting and giving their clinical data of this experimental analysis. Further, we are grateful to Dr. Challa Vamsi Krishna and his supporting staff for extending their cooperation and suggestions towards this research work.

REFERENCES

1. Feng, B., Zhao, Y. Y., Wang, J., Yu, H., Potu, S., Wang, J& Guo, Y. (2019, May). "The Application of Machine Learning Algorithms to Diagnose CKD Stages and Identify Critical Metabolites Features". In International Work-Conference on Bioinformatics and Biomedical Engineering (pp. 72-83). Springer, Cham. https:// doi.org/ 10.1007/ 978-3-030-17938-0_7
2. Abdelaziz, A., Salama, A. S., Riad, A. M., & Mahmoud, A. N. (2019). "A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities". In Security in Smart Cities: Models, Applications, and Challenges (pp. 93-114). Springer, Cham. https://doi.org/10.1007/978-3-030- 01560-2_5
3. Besra, B., & Majhi, B. (2019). "An Analysis on Chronic Kidney Disease Prediction System: Cleaning, Preprocessing, and Effective Classification of Data". In Recent Findings in Intelligent Computing Techniques (pp.473-480) Springer, Singapore.https://doi.org/10.1007/978-981-10-8639-7_49
4. Hasan, K. Z., & Hasan, M. Z. (2019). Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction o"f Chronic Kidney Disease". In Emerging Research in Computing, Information, Communication and Applications(pp. 415-426). Springer, Singapore.https://doi.org/10.1007/978-981-13-5953-8_34
5. Hore, S., Chatterjee, S., Shaw, R. K., Dey, N., & Virmani, J. (2018). "Detection of chronic kidney disease: A NN-GA-based approach". In Nature Inspired Computing (pp. 109-115). Springer, Singapore. https://doi.org/10.1007/978-981-10-6747-1_13
6. Alaoui, S. S., Aksasse, B., & Farhaoui, Y. (2018, July). "Statistical and Predictive Analytics of Chronic Kidney Disease". In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 27-38). Springer, Cham.shttps://doi.org/10.1007/978-3-030-11884-6_3
7. Kriplani, H., Patel, B., & Roy, S. (2019). "Prediction of Chronic Kidney Diseases Using Deep Artificial Neural Network Technique". In Computer Aided Intervention and Diagnostics in Clinical and Medical Images (pp. 179-187). Springer, Cham. https://doi.org/10.1007/978-3-030-04061-1_18
8. Chatterjee, S., Dzitac, S., Sen, S., Rohatinovici, N. C., Dey, N., Ashour, A. S., & Balas, V. E. (2017, June). "Hybrid modified Cuckoo Search-Neural Network in chronic kidney disease classification". In 2017 14th International Conference on

Engineering of Modern Electric Systems (EMES) (pp. 164-167). IEEE. 10.1109/EMES.2017.7980405
9. Salekin, A., & Stankovic, J. (2016, October). "Detection of chronic kidney disease and selecting important predictive attributes". In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 262-270). IEEE. 10.1109/ICHI.2016.36
10. Subasi, A., Alickovic, E., & Kevric, J. (2017). "Diagnosis of chronic kidney disease by using random forest". In CMBEBIH 2017 (pp. 589-594).Springer,Singapore. https://doi.org/10.1007/978-981-10-4166-2_ 89
11. Boukenze, B., Haqiq, A., & Mousannif, H. (2016, May). "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques". In International Symposium on Ubiquitous Networking (pp. 701-712). Springer, Singapore. https://doi.org/10.1007/978-981-10-1627-1_55
12. Ravindra, B. V., Sriraam, N., & Geetha, M. (2018). "Classification of non-chronic and chronic kidney disease using SVM neural networks." International Journal of Engineering & Technology, 7(1.3), 191-194.
13. Chaitanya, S. M. K., & Kumar, P. R. (2019)." Detection of Chronic Kidney Disease by Using Artificial Neural Networks and Gravitational Search Algorithm". In Innovations in Electronics and Communication Engineering (pp. 441-448). Springer, Singapore. https://doi.org/10.1007/978-981-10-8204-7_44
14. Ruggieri, Salvatore. "Efficient C4. 5 [classification algorithm]." IEEE transactions on knowledge and data engineering 14, no. 2 (2002): 438-444
15. Vital, T. Panduranga, GSV Prasada Raju, IS Siva Rao, and AD Praveen Kumar. "Data collection, statistical analysis and machine learning studies of cancer dataset from north costal districts of AP, India." Procedia Computer Science 48 (2015): 706-714.
16. Gadde, Praveen, Suresh Sanikommu, Ramesh Manumanthu, and Anitha Akkaloori. "Uddanam nephropathy in India: a challenge for epidemiologists." Bulletin of the World Health Organization 95, no. 12 (2017): 848.
17. Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology, 4(12), 608-612
18. Dhayanand,& Vijayarani, S., S. (2015). Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics & Informatics (IJCI), 4(4), 13-25.
19. Ani, R., Sasi, G., Sankar, U. R., & Deepa, O. S. (2016, September). Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1287-1292). IEEE.
20. Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. Indian Journal of Science and Technology, 9(29), 1-6.
21. Borisagar, N., Barad, D., & Raval, P. (2017). Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm. In Proceedings of International Conference on Communication and Networks (pp. 295-303). Springer, Singapore.
22. Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of medical systems, 41(4), 55.
23. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2).
24. Baby, P.S. and Vital, T.P., 2015. Statistical analysis and predicting kidney diseases using machine learning algorithms. International Journal of Engineering Research and Technology, 4(7).

AUTHORS PROFILE

Smt. K. B. Anusha pursued B.Tech in Information Technology from JNTUK of A.P., She has 7 years of teaching experience in Aditya Institute of Technology and Management (AITAM), Tekkali. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Aditya Institute of Technology and Management (AITAM), India. She is a member of ACM, CSI and IMRTC She has published 5 papers in reputed international journals and 1 in conference. Her main research interests are Mobile Computing, Computer Networks, Machine Learning, Deep Learning and IoT.

Dr. PanduRanga Vital Terlapu pursed Bachelor of Science in Computer Science from Andhra University of A.P, India in 1995 and Master of computer Application from Andhra University in year 1998. He completed his M. Tech in Computer Science and Engineering from Acharya Nagarjuna University of A.P, India and he completed his Ph.D in Computer Science and Engineering from GITAM University of A.P, India. He has 19 years of teaching and 13 years of research experience. He is currently working as Associate Professor in Department of Computer Science and Engineering, Aditya Institute of Technology and Management (AITAM), India. He is a member of ACM, Life Time Membership from International Computer Science and Engineering Society (ICSES), USA and Life Time Membership from Indian Society for Technical Education (ISTE), New Delhi, India. He has published more than 30 research papers in reputed international journals including SCOPUS indexed and a conference including Sringer, Elsevier and it's also available online. He is reviewer of reputed journals like Springer, Elsevier and IEEE. His main research work focuses on Machine Learning, Deep Learning and Data Mining, Data and Big Data Analytics, IoT and Computational Intelligence, Voice Analysis and Voice Processing, Bioinformatics.

Smt. kurman Sangeeta, Sr.Assistant Professor,CSE Department ,Aitam Tekkali.She did her M.Tech in Computer science and Engineering from IIT Madras. Her research interest includes machine Learning,Artificial Intelligence,Deep Learning and Cryptography. She has 13 yrs of experience in teaching field and one year in industrial area. She has 6 publications in International journals,2 in national journals and presented 2 papers in national and 3 papers in international conferences