A NEURAL NETWORK APPROACH FOR CLASSIFICATION OF KIDNEY DISEASE DATASETS COLLECTED FROM VISAKHAPATNAM OF A.P., INDIA

T. Panduranga Vital[1], M.Y.V. Nagesh[2], T. Anuradha[2],
Appala Raju Samanthula[2]

[1]Department of CSE, Scholar of GITAM University, Rushikonda, Visakhapatnam, (India)
[2]Department of CSE, Scholar of GITAM University,
Raghu Engineering College, Visakhapatnam, (India)

ABSTRACT

Data mining is a vital tool to learn the data for Diseased datasets which deal with huge amount of data. This paper reveals the predicting Kidney Disease by using Neural Network Approach. The data has been collected from Visakhapatnam District of AP, India during the years 2014 -2015 with 1380 instances (kidney Diseased 690 instances and 690 Healthy instances) with 50 attributes. Here we are taking mainly health check profiles such as age, height, weight, gender, blood pressure, blood sugar, water intake , calcium(food habits and tablets), monthly enrolls and so on, it can forecast the patients getting a kidney disease or not. The methods like feed forward, probabilistic neural networks(PNN) that confusion matrix, Self Organization Maps (SOM) for unsupervised clustring and for prospect predition- dynamic time series were analysed using Mat Lab in Neural Networks. The Results are shown in each method has its unique strong point in the definite mining goals. The dataset is justified as 100% correct classification for prediction of kidney disease in PNN.

Keywords: Data mining; Kidney, Visakhapatnam, neural networks, PNN

I. INTRODUCTION

Data mining is the headway of extracting the patterns from data and convert to useful in sequence. Data mining is becoming vital tool in the current decades to make over the data into information. Study of Kidney Diseased datasets is one of the most important researches in data mining techniques. Data mining software's like Mat Lab has a number of diagnostic tools used for analyze in order from particular outlooks like machine learning and summarizing the data into useful information [14].The analysis and summarization can be used to raise accuracy of the data. Researchers in many fields like business management, computer sciences, biology and communication networks have shown great curiosity in data mining.[15]

Standard multilayer feed forward networks has one hidden layer using random functions are competent of approximating any Borel measurable function from finite dimensional space to a further to any preferred degree of accuracy, provided many hidden layers are available. In this sense, multilayer feed-forward networks are a class of collective approximations. [2] The effectiveness of the proposed greatest likelihood training algorithm is assessed using nonparametric arithmetical methods to define acceptance interval on PNN classification performance [7]. The ability of radial basis function networks and kernel neural networks that a exact probabilistic neural networks .And study their similarities and difference. In order to pass up the huge amount of hidden unit of the kernel neural networks or probabilistic neural networks and decrease the training time for the radial basis function networks [3]

Many neural network classifiers afford outputs which approximation Bayesian a posteriori probabilities. When the estimation is correct, network outputs can be accurate that Assessment accuracy depends on network difficulty. The sum of training data, and the sum to which training data replicate true probability distributions and a priori class probabilities, accepting of network outputs as Bayesian probabilities allow outputs from multiple networks to be collective for higher level conclusion, make things easier creation of rejection thresholds, make it possible to pay damages for difference between pattern class probabilities in training and test data, it allows outputs to be diminish different risk functions, and suggest alternative events of network performance.[5]

A probabilistic neural network can calculate nonlinear decision limits which approach the Bayes finest is formed. A four-layer neural network projected can map any input sample to any number of classifications [1]. Computationally, the back-propagation neural network is at a solemn complexity to maximum probability, captivating nearly an order of size more computing time when implied on a serial workstation.[8]

Two divergent neural-network approaches namely PNN and self-organized map (SOM) were examined, and their recitations were benchmarked. Adult groups are at a higher possibility of mortality associated with air pollution. Such complication should be taken into relation in health risk estimation based on time series studies.[16] Researcher classically study of time series cross section data with a binary needy variable is common. The number of such observations appears to be rising exponentially Patients with chronic kidney disease lose kidney function and in overindulgence of time are at threat of rising end stage kidney disease. Predict individuals at risk for chronic kidney disease is significant first step in modify the progressive course of chronic kidney disease. Early recognition of chronic kidney disease would afford the best opportunity to apply strategies known to brake the loss of kidney function [12]. People with kidney disease have a risk of coronary actions like to those with preceding myocardial infarction. We assess whether chronic kidney disease be supposed to observe as a coronary heart disease threat equivalent [11].

Biomedical research applies a extensive choice of designs from different questionnaires from patients to solve tribulations in clinical laboratories and residents settings.

II. METHODOLOGY

The main aim of processing the data in the current experimentation is to discriminate healthy people from those with kidney disease with a two-decision classification problem. Mat lab version 7.12.0.635 (R2011a) was applied in the present experimentation for analysis of neural network approach.

The data has been collected from Visakhapatnam district of AP, India during the year 2014-2015 with 1380 instances(kidney Diseased 690 and Healthy 690) with 50 attributes (Gender, Age, Height, Weight, Blood Group, Body-Color, Job-Position, Place, Food-Habits, Meals-Regularity, Breakfast-Items, Lunch-Items, Dinner-

Items, Non-veg.--Number-of-Times-A-week, Salt-Consumption, Prefer-Fruits, Kinds Of Fruits Preferred, Prefer-Leafy-Vegetables, Kinds Of Leafy-Vegetables-Preferred, Fast-foods-Preference, Soft-drinks-Intake, Prefer-Tea, No-Of-Times-Tea, Prefer-Coffee, No-Of-Times-Coffee, Prefer-Milk/Milk-product, No-Of-Times-Milk, Smoking Habit, No-Of-Times-Smoking, Drinking Habit, No-Of-Times-Drink, Water Intake, Type-Of-Water, Type-Of-Soil, Other Diseases, HighlyUsedTablet1, HighlyUsedTablet2, HighlyUsedTablet3, History Of Kidney stones, RelationToMember, Sweat Formation, Pregnant, Any Surgeries Before, Yoga/Meditation, Regular Vomiting Sensation, Body Temperature In A Week, Kidney stone Before, Bathing, Sleeping Hours)

III. RESULTS AND DICUSIONS

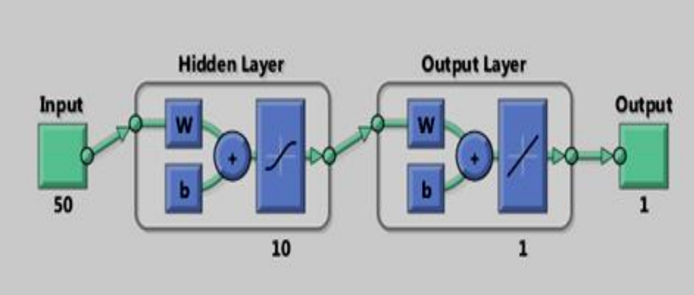
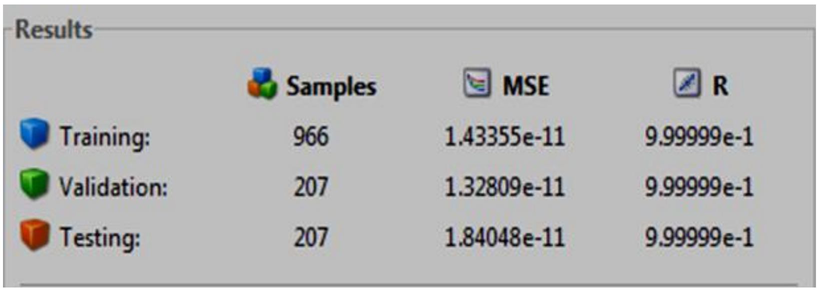


Fig.-1: Feed – Forward network

Fig.-1 shows the two layered Feed – Forward network with 50 inputs 10 sigmoid hidden neurons and linear fit net (output) neuron. The inputs ‘Kidney Diseased’ is a 50X1380 matrix representing the static data 1380 (and samples of 50 elements. The targets ‘Kidney Target’ is 1X1380 matrix (0 for non-disease 1 for diseased).The whole dataset is divided into three kinds of samples i.e. Training is 966(70%)) validation is 15% and Testing is 15%.



Results

	Samples	MSE	R
Training:	966	1.43355e-11	9.99999e-1
Validation:	207	1.32809e-11	9.99999e-1
Testing:	207	1.84048e-11	9.99999e-1

Fig.-2:Best Performance Samples, MSE and R Values

Fig.-2 shows the best performance of each dataset samples. The Mean Squared Error nearer to zero for all Training , Validation and testing. Regression R value is nearer to one. So data validation is 100% correct classification.

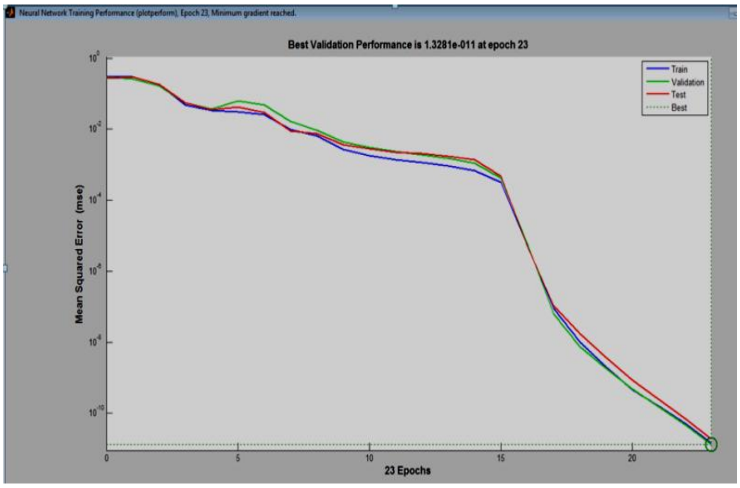


Fig.-3: Training Performance of Kidney dataset

Fig.-3shows Training Performance, The Best validation performance is 1.3281e-011 at epoch 23.

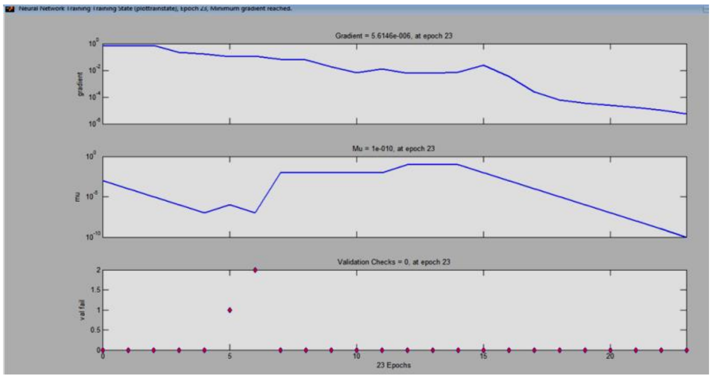


Fig.-4: Plot Training State

Fig.-4 shows the plot training state shows the values gradient 5.614e-006 at epoch 23,Mu =1 at epoch 23 and validation checks =0 at epoch 23. Fig.-5 shows the Training Error Histogram with 20 bins. Zero error nearer at -1e-007.

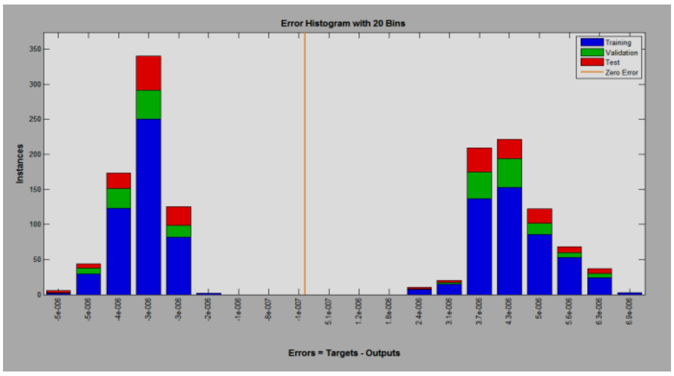
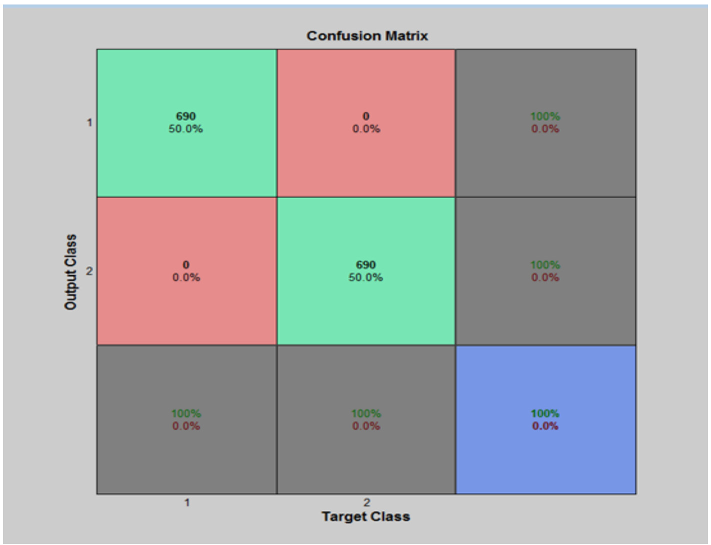


Fig.-5: Training Error Histogram with 20 Bins

To analyze the probabilistic neural network response, confusion matrix is computed by taking into consideration the outputs of the trained network and comparing with the expected target results shown in Fig. 6.



Confusion Matrix

Output Class	Target Class 1	Target Class 2	
1	690 50.0%	0 0.0%	100% 0.0%
2	0 0.0%	690 50.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%

Fig.-6: Confusion Matrix

The crosswise cells verify the number of true sets that were fittingly classified for each class of kidney patients. The off diagonal cells gave the details that the number of remains positions that were misclassified. The subsequent results presents the precision obtained by training the probabilistic neural network using Kidney dataset and got 100% of data for training as positives (correctly classified) using Mat lab.

Fig.-7 describes the Receiver Operating Characteristic curve (ROC) The blue colored plotting line shown in Fig-7 describes the ROC plotted in 2-dimentional between sensitivity(the false positive rate) and specificity(the false positive rate). The ROC value is 1.So the Kidney Dataset is 100% accurate to predict the kidney disease patients.

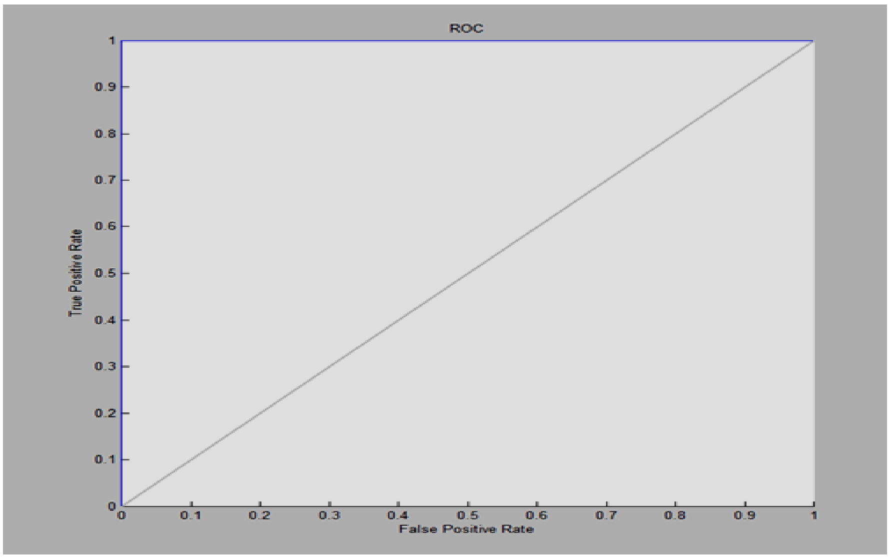


Fig.-7: Receiver Operating Characteristic Curve

Fig.-8 was shown best validation performance as 1.2819e-077 at epoch 26.

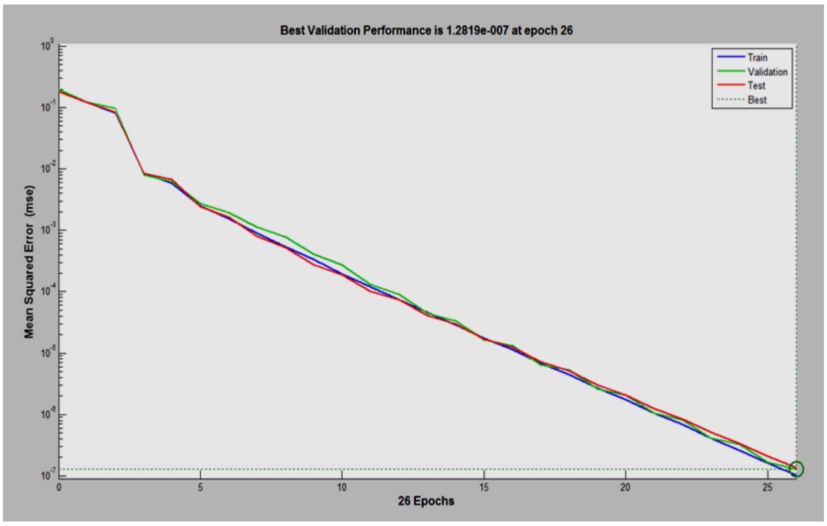


Fig.-8: Best Validation Performance

Fig.-9 shows Self organizing Map neural cluster tool architecture. In this the neurons are arranged in a two-dimensional topology as well two-dimensional approximation. With using 50 inputs it creates the 10X10 orthogonal topology.

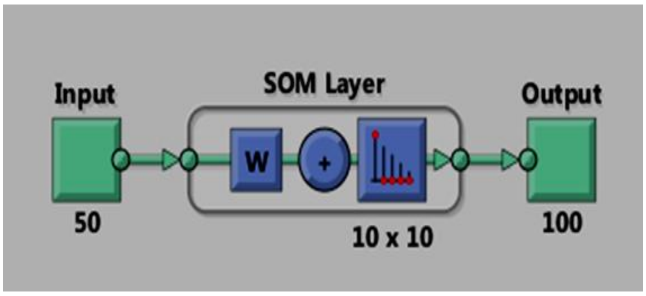


Fig.-9: SOM Neural Network

Fig.-10 shows the SOM Neighbor Distances. SOM network clustered into two different groups. The darker colors tell the larger distances vice versa lighter tells smaller distances.

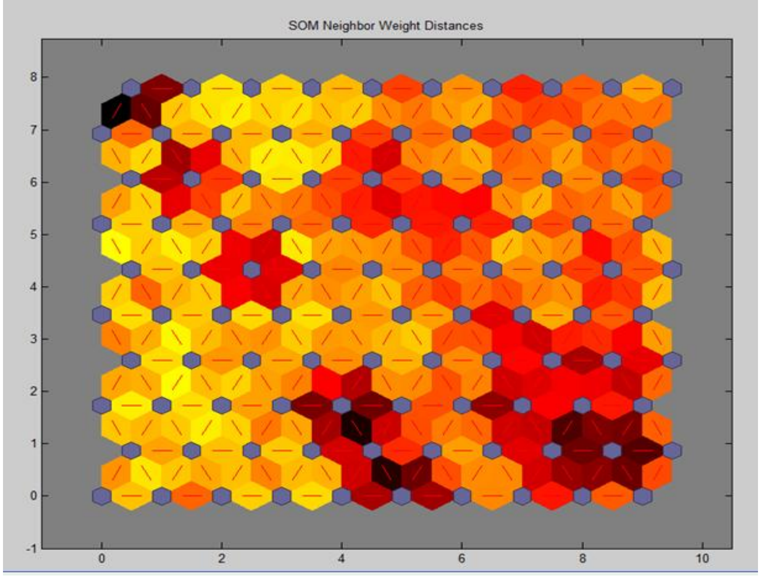


Fig.-10: SOM Neighbor Distances

Fig.-11 shows the SOM Weight position.The Weight 1 values 0 to 1. The weight2 values are between 0 to 120.

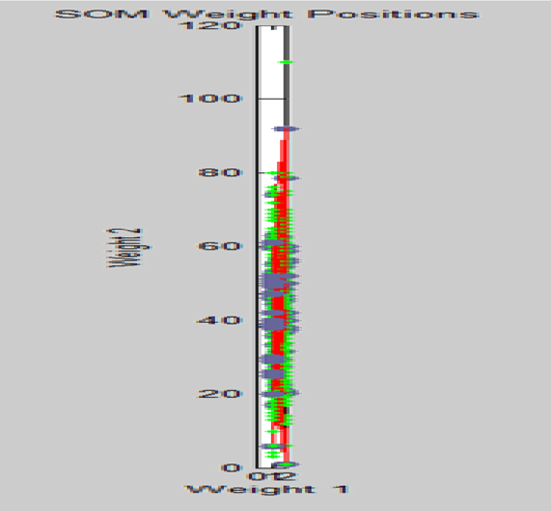


Fig.-11 SOM- Weight Positions

Fig.-12 shows the SOM number of hits linked with each neuron. The SOM is hexagonal. It indicates the training data count is associated with each of the neuron. The topology is a 10-by-10 network then there are 100 neurons, maximum number of hits associated with that neuron is 32 and Minimum number of hits associated with that neuron is 1.

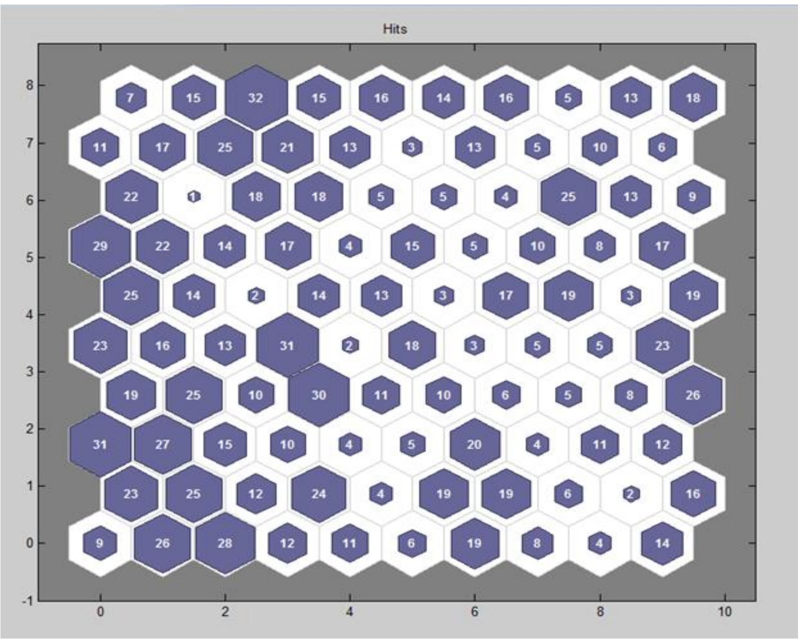


Fig.-12: SOM Number of Hits Associated with Each Neuron

Fig.-13 shows the Dynamic Time series Neural Network model for the kidney dataset for monthly enrolments.

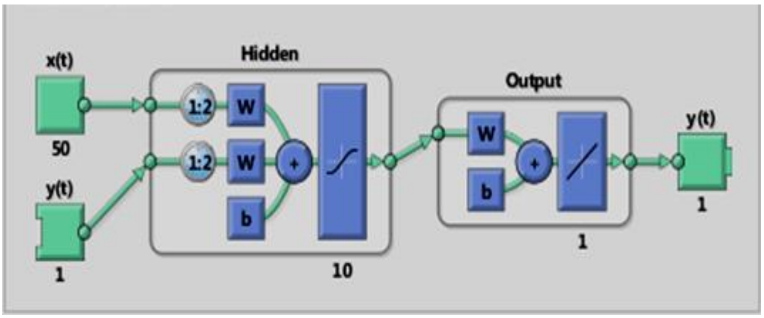
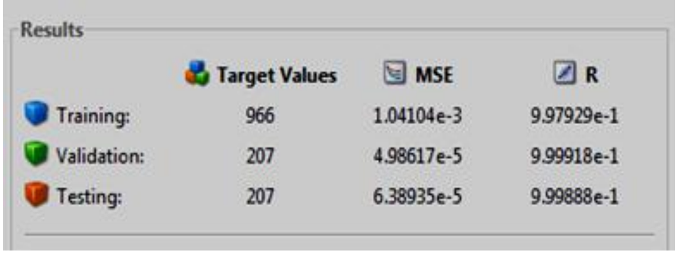


Fig.-13: Dynamic Time series Neural Network Model

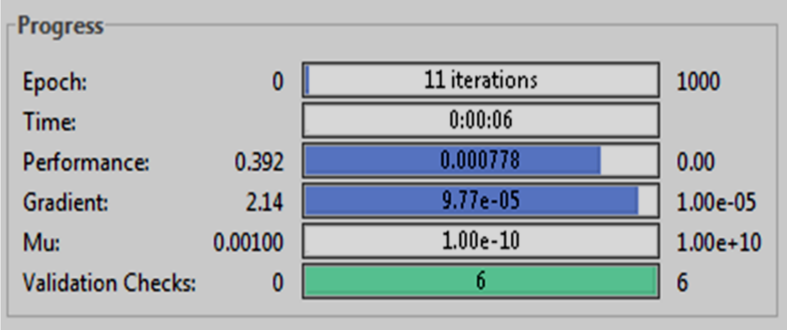
Fig-14 shows the Best Performance that Target Values, MSE and R Values for the training, validation and testing respectively. The good results of R values are nearer to 0.99 to all the Training, Validation and Testing. So the performance of the Kidney dataset is good.



Results

	Target Values	MSE	R
Training:	966	1.04104e-3	9.97929e-1
Validation:	207	4.98617e-5	9.99918e-1
Testing:	207	6.38935e-5	9.99888e-1

Fig.-14 Best Performance Target Values, MSE and R Values



Progress

Epoch:	0	11 iterations	1000
Time:		0:00:06	
Performance:	0.392	0.000778	0.00
Gradient:	2.14	9.77e-05	1.00e-05
Mu:	0.00100	1.00e-10	1.00e+10
Validation Checks:	0	6	6

Fig.-15: Dynamic Time Series NN Progress

Fig-15 shows the Progress of the Dynamic time series Network that classifies Kidney dataset in 11 iterations , takes time 00:06 sec and 6 validation checks.

Fig-16 describes the plot performance at Epoch 11. The best Validation performance is 0.00011612 at epoch 5.

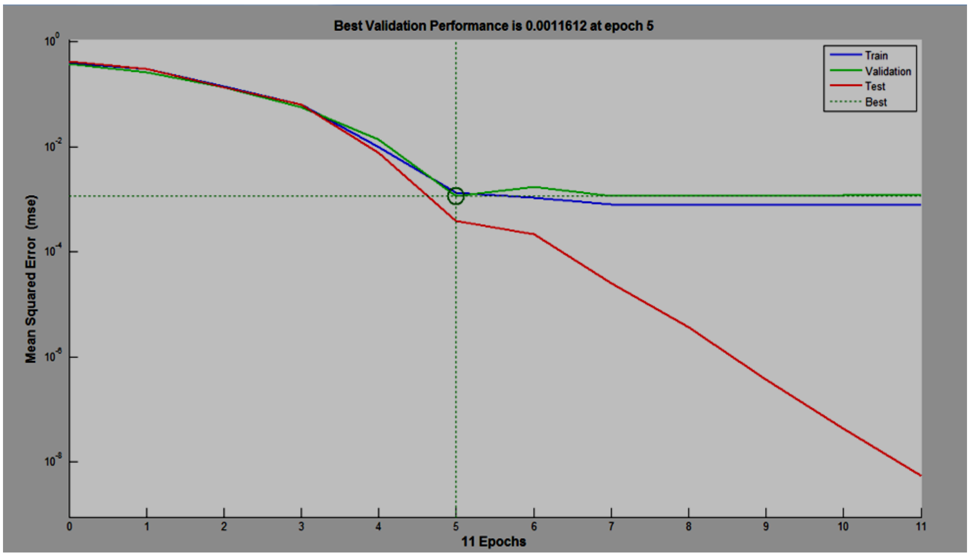


Fig-16 Neural Network Training performance Epoch 11

Fig-17 shows the Training States of the Time series Neural Network Gradient value is 9.7668e-005, the Mu value is 1e-010 and 6 Validation checks at epoch 11.

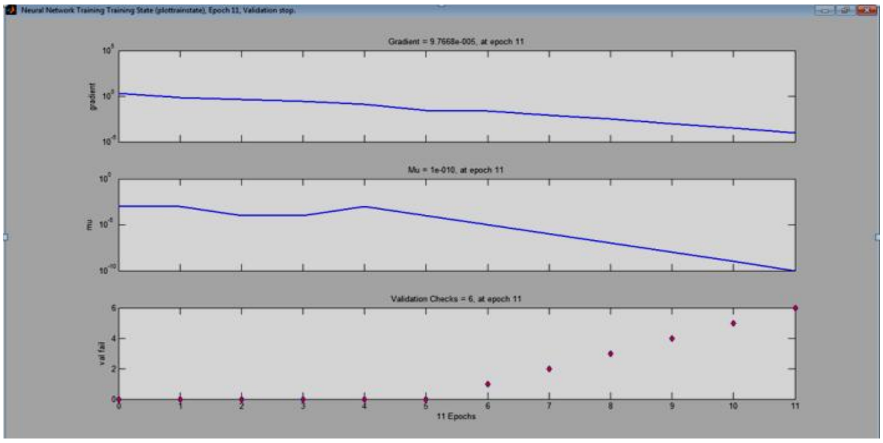

Neural Network Training Training State (plottrainstate), Epoch 11, Validation stop.
Gradient = 9.7668e-005, at epoch 11
gradient
10^{5}
10^{0}
10^{-5}
Mu = 1e-010, at epoch 11
mu
10^{0}
10^{-5}
10^{-10}
Validation Checks = 6, at epoch 11
val fail
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
11
11 Epochs

Fig.-17: Neural Network Training States

Fig-18 describes the R (Regression) values of Training. Validation and Testing. The Training R value is 0.99772. The Validation R value is 0.99963. And the Testing R value is 0.99818.All Regrssion(R) value is 0.99808.So the kidney Dataset is good for future predictable values.

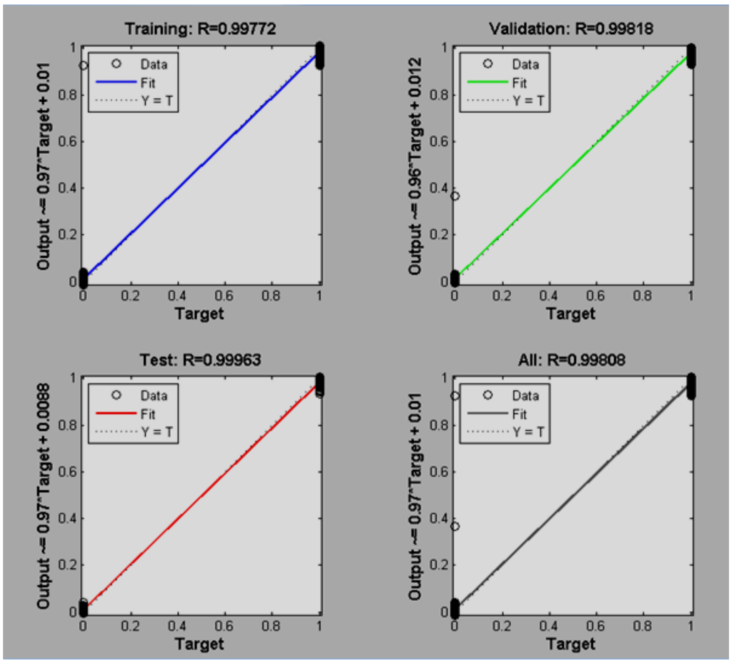


Fig.-18: Regression (R) Values

IV. CONCLUSION

Classification, clustering and association are significant among the techniques of data mining. Accuracy is the main goal to estimate the performance of the algorithms over kidney disease dataset. The PNN method using Matlab was provide data justification as 100% correct classification. The probability advance was provided good results for KIDNEY datasets and supplementary analysis can give kind disease for diagnosis and cure.

V. ACKNOWLEDGMENT

The authors would like to thank GITAM University and Raghu Engineering College for provided that computational competence and access to e-journals to bear out this research.

REFERENCES

[1] Specht, D. F. (1990). Probabilistic neural networks. Neural networks, 3(1), 109-118.

[2] Hornik et al., (1989). Multilayer feed forward networks are universal approximators. Neural networks, 2(5), 359-366.

[3] Huang, D. S. (1999). Radial basis probabilistic neural networks: Model and application. International Journal of Pattern Recognition and Artificial Intelligence, 13(07), 1083-1101.

[4] Temurtas et al., (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications, 36(4), 8610-8615.[8]

[5] Swapna, et al., (2013). Automated detection of diabetes using higher order spectral features extracted from heart rate signals. Intelligent Data Analysis, 17(2), 309-326.

[6] Lima, A. et al., (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. Neural Networks, 58, 122-130.

[7] Streit et al., (1994). Maximum likelihood training of probabilistic neural networks. Neural Networks, IEEE Transactions on, 5(5), 764-783.

[8] Paola, J. D et al. (1995) A detailed comparison of back propagation neural network and maximum-likelihood classifiers for urban land use classification. Geoscience and Remote Sensing, IEEE Transactions on, 33(4), 981-996.

[9] Tian et al., (1999). A study of cloud classification with neural networks using spectral and textural features. Neural Networks, IEEE Transactions on, 10(1), 138-151.

[10] Lawrence, et al., (1998). Neural network classification and prior class probabilities. In Neural networks: tricks of the trade (pp. 299-313). Springer Berlin Heidelberg.

[11] Tonelli et al., Disease Network. (2012). Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. The Lancet, 380(9844), 807-814.

[12] Kshirsagar et al., (2008) A simple algorithm to predict incident kidney disease. Archives of internal medicine, 168(22) 2466-2473.

[13] Camps-Valls, et al., (2003) Prediction of cyclosporine dosage in patients after kidney transplantation using neural networks. Biomedical Engineering, IEEE Transactions on 50(4), 442-448.

[14] H. C. Wu, and C. N. Lu, “A data mining approach for spatial modeling in small area load forecast,” Power Systems, IEEE Transactions on, vol. 17, no. 2, pp. 516-521, 2002.

[15] M. Goebel, et al. “A survey of data mining and knowledge discovery software tools, “ ACM SIGKDD Explorations Newsletter, vol. 1, no. 1, pp. 20-33, 1999.

[16] Gouveia, N., & Fletcher, T. (2000). Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status. Journal of epidemiology and community health, 54(10), 750-755.

[17] Beck, N., Katz, J. N., & Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. American Journal of Political Science, 1260-1288.

[18] Reis, B. Y., & Mandl, K. D. (2003). Time series modeling for syndromic surveillance. BMC Medical Informatics and Decision Making, 3(1), 2.