PREDICTION OF KIDNEY FAILURE IN BIG DATA

Mrs S.Ancy
Assistant Proffessor'
Department of IT,
JeppiaarInstituteof
Technology,
Chennai,india,
sancy@jeppiaarinstitute.org

Miss H.Keerthana
Final year
Department of IT,
JeppiaarInstituteof
Technology,
Chennai,India,
Keerthanakiran2326@gmail.com

Miss.S.Preethi
Final year
Department of IT,
JeppiaarInstituteof
Technology,
Chennai,India
preethiselvaraj14@gmail.com

Miss.S.Sharmila
Final year
Department of IT
JeppiaarInstituteof
Technology,
Chennai,India,
sharrmisridharan@gmail.com

Abstract— **Big data is the large set of dataset used for extraction , selection , analyzing and interpolation of data. Big data is used wide assortment in medical fields for analyzing the patients , prediction of future effects and clinical decision making it is a tool used to store large number of datas to understand the diseases and to predict the disease in future In this paper we use RBFNN (Radial Basis Function Neural Network) with classifier algorithm with the use of parameters to determine the condition of a patient as a normal or a kidney failure patient**
The proposed method reveals the stages of the kidney failure patient and treatment and clinical decision

Keywords— *big data , Radial basis function neural network ,kidney failure, classifier algorithm , clinical decision*

I. Introduction

Big Data may well be the Next Big Thing in the IT world. Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis . Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing technique.The characteristics of big data are volume ,velocity and variety .Big data is different from traditional computing Automatically generated by a machine , Typically an entirely new source of data , Not designed to be friendly , May not have much values. It is used to Examine large amount of data ,Appropriate information , Identification of hidden patterns, unknown correlations , Competitive advantage , Better business decisions: strategic and operational , Effective marketing, customer satisfaction, increased revenue we can take data from any source and analyze it to enable with the features of cost reductions, time reductions, new product development and optimized offerings and smart decision making.

Big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods [15] , big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions [2-15] and as a research category. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. big data analytics has the potential to improve care, save lives and lower costs. Big data analytics is to proactively identify individuals who would benefit from preventative care or lifestyle changes; broad scale disease profiling to identify predictive events and support prevention initiatives; collecting and publishing data on medical procedures,

Kidney failure is the silent death and non communicable disease. Physicists , chemists and clinicians to predict function of a kidney that would give doctors a better chance to intervene to save a patient and improve patient recovery periods It is a bit like measuring GDP in the economy: a single number quantifying a huge amount of complex activity and allowing you to understand the dynamics of the system. Development and validation of prediction models using demographic, clinical, and laboratory data of the patients. It is used to utilize clinical and nonclinical data sources to develop predictive models (PMs) that identify patients with a high probability of not This was translated to a single number (on a continuous scale from 0 to 1) describing the likelihood of a patient's state at any one time resulting in organ success or failure.

Big data affects organizations across practically every industry.

- Manufacturing- The manufactures can check the quality and can show the output.it will minimize the waste that has been occurred .many of the manufactures will work on analytics-based culture. It can solve the problems and make work faster .the manufactures can store the details of customer and clients

- Government- big data can help the government to store the details and update the information of the sectors .this helps government to check and prevent from crime. The main advantage of big data is government must address the issues of transparency and privacy
- Banking-bank can store large amount of information of the customer details. Where it can be change,update and view the details . Each and every transaction will be stored in big data it will boost the satisfaction of customer
- Retail- the customer relationship is more critical in retail industry. Big data is one of the best way to manage the details of the customer .they will have clear view about the customer requirements. Big data will also help their transaction and update their details
- Fraud management- it will detect the fraud and improve the prevention .in real time system storing the information can be easily detected by fraud .but in big data storing information cannot be detected by anyone it increases the productivity. Using big data is risk been reduced
- Demand signal management-In demand driven organization with demand signal management software can be used in some platforms. It is used to capture the market and retailer data and combine it with business .main advantage is demand signals will be more faster

II. LITERATURE SURVEY

1. Navdeep told about " A Predictive Progression Of Chronic Kidney Diesease To Kidney Failure"

Chronic Kidney Diesease is Common This Type of kidney disease can be classified by Estimated Glomerular FilterationRate(GFR) and Albuminuria . It deals About to Develop and Validate Predictive Models for The Progression of CKD. The accurate model included Sex,Estimated GFR, Albumaniria , Serum Bicarbonate, Serum Calcium and Serum Phosphate which produces a accurate information than considering Albumin, Sex and GFR . It can give accurate information The main disadvantage is that the GFR is not alone sufficient for clinical decision making

2.L.KONONENKO,"Machine learning for medical diagnosis: history, state of art and perspective,"Artifical Intelligence in medicine" vol.23, no.1, pp.89-109,2001

The development of intelligent data analysis in medicine from a machine learning perspective. for applying machine learning in medical diagnosis. In this paper the algorithm used are the naive Bayesian classifier, neural networks and decision trees. . The first describes a recently developed method for dealing with reliability of decisions of classifiers, which seems to be promising for intelligent data analysis in medicine. and it is used approach to using machine learning in order to verify some unexplained phenomena from complementary medicine, in the future medical diagnosis and treatment play an important role in orthodox community

3.A.AZEMI,V.R.SABZEVARI,M.KHADEMI,H.GHOLIZADE,A.KIANI,Z.S.Dastgherib"INTELLIGENT ARRHYTHMIA DETECTION AND CLASSIFICATION USING ICA" in engineering in medicine and biology society,2006 EMBS 28th Annual International conference of IEEE,2006,pp.2163-2166

Knowledge-based method for arrhythmic beat classification and arrhythmic episode detection and classification using only the RR-interval signal extracted from ECG recordings. A three RR-interval sliding window is used in arrhythmic beat classification algorithm. four categories of beats: normal, premature ventricular contractions, ventricular flutter/fibrillation and 2 degrees heart block. The beat classification is used as input of a knowledge-based deterministic automaton to achieve arrhythmic episode detection and couplet, ventricular tachycardia, ventricular flutter/fibrillation and 2 degrees heart block. The method is evaluated by using the MIT-BIH arrhythmia database. It gives a high performance of 98% accuracy for arrhythmic beat classification and 94% accuracy for arrhythmic episode detection and classification. The only limitations is it uses only the RR-interval signal for arrhythmia beat and episode classification

.

4. H. L. Lu, KONG and P.CHIA "An automated ECG classification system based on neuro - fuzzy system", in computer in Cardiology 2000, 2000, pp.387-390

12-lead electrocardiogram (ECG), in the early diagnosis of acute myocardial infarction (AMI)., a hybrid neuro-fuzzy approach to the diagnostic classification of 12-lead ECGs is presented. The architecture used is a combination of fuzzy logic and neural network theory. it benefits from the reasoning capabilities of fuzzy logic as well as the learning ability of neural networks. It consists of two phases: (1) Use fuzzy logic to establish the diagnosis system, (2) it uses a back propagation network to automatically adjust the parameters of the system. A total of 124 ECGs from patients with or without acute myocardial infarction. Sensitivity, specificity, partial and total accuracy are used for evaluation of the system. The neuro-fuzzy system correctly identified 89.4% of the patients with AMI and 95.0% of the patients without AMI. The AMI can be diagnosed with reasonable accuracy but it consumes time and the diagnosis of AMI varies according to clinical circumstances

5. T.POGGIE AND F.GIROSI,"Networks for approximation and learning "proceeding of the IEEE, vol.78, pp.1481-1497, 1990
The problem of the approximation of nonlinear mapping, (especially continuous mappings) i Regularization theory and a theoretical framework for approximation (based on regularization techniques) that leads to a class of three-layer networks called regularization networks it related to the radial basis functions, mainly used for strict interpolation tasks. The extensions of regularization approach to splines , regularization, Bayes formulation, and clustering. The regularization networks are generalized to a formulation that includes task-dependent clustering and dimensionality reduction. Theoretical framework for approximation based on regularization techniques that leads to a class of three-layer networks Radial Basis Functions, mainly used for strict interpolation tasks. But it also related pattern recognition methods. The main disadvantage of an exact representation of the approximation of linear and nonlinear mappings in terms of simpler functions of fewer variables.

6. Kayyan Najarian, Daniel A. Beard in "European Journal of scientific research" told about "the Prediction Of Kidney Failure Using Artificial method Neural Network" in 2011
Many persons are affected due to Kidney failure this End stage Kidney Disease and require some form of Renal Replacement Theraphy to sustain life. More than 50 Parameters may be monitored while providing a kidney Dialysis Treatment. Prediction Of Kidney Failure using artificial Neural Networks Explored and analyzed The massive data generated for several years from the kidney patients using Artificial neural Network To Select The Input Variables ,Weight and connectivity structure to determine Optimal Network Topology and To Train The System for Learning . This helps to determine the Features that are predictive of a patient life Expectancy, detect the existance of Renal Failure in Patient and to predict the kidney dialysis survival. The advantage is it can automatically assess the dialysis service. The Main disadvantage in this paper is there are 50 parameters take more time and multiple assumption makes the doctor in confusion to predict the kidney failure

7. Gerardo in "Cardiac Sudden Death Risk Detection Using Hybrid Neuronal Fuzzy Networks"
The Cardiovascular Diesease are the Main Cause of Morality. A Non Invasive technique is the analysis of HRV (Heart Rate Variability) often from , ECG record of 24 hours . HRV is the measurement of the interval between R peaks of two Consecutive QRS complex .An Adaptive filter is used to Eliminate the Noise Signals , Finally Power Spectral Density are computed,Filtering the signals in three band characterize the HRV: High Frequencies(HRV),Low Frequencies(LRV) and Very Low Frequencies (VLF).The inputs are from the Time and Frequency Domain. This Algorithm helps in Quantification and characterization of HRV. This helps the Patients in identifying high/low Cardiac Risk. The Main Disadvantage is It does not detect the accurate and approximation values And Doctors cannot determine suddenly

8.Shruthi Told about " A Survey on RBF Neural Network for intrusion Detection System"
Network security is a Hot Burning issue . With the help of technology Advancement Hackers are adopting new methods to create different attacks in order to harm network security . IDS is a kind of Security software which inspects all incoming and outgoing traffic . It will generate any unusual behavior is found in a network . IDS uses a various approaches such as Data mining,Neural network and Statistical approach . Neural network is the best approach among these RBF is a feed forward and supervise technique of Neural Network. It is used for anamoly Detection and Misuse Detection . The main disadvantage is that We need to do proper optimization of RBF parameters and also response time will increase. Instead of using Sequential Layered Approach for IDS we shall use Parrallel Layered Approach

9. Messadag told about " An approach for ECG using Wavelets classification and Markov model**"**
Cardiac attacks can be easily monitored using ECG . It say about the relaxation and contraction of heart it gives a graphic and democrartic view of a patient heart. This paper deals about the classification of ECG signal. This method is done using the Ergodic Markov Model.The data was obtained from MIT BIH arrhythmia in two categories .1) Each beat is isolated and its discrete wavelet transform is calculated 2) vectotial Quantization is applied. This is based on a Non linearTransformation . The parameters can be taken as a P,Q,R and s. and then the ECG can be calculated. These parameters of Markov model are computed for different number of states. This paper Disadvantage is it cannot be easily predict and finalized with ECG alone and time gets consumed

10.NicosMaglaveras Told About " ECG Pattern Recognition and classification using Linear transformation and Non linear transformation**"**
The most widely used signal in clinical practice is the ECG. ECG conveys information regarding the electrical function of the heart, by altering the shape of its constituent waves,

namely the P, QRS, and T waves. Thus, the required tasks of ECG processing are the reliable recognition of these waves, and the accurate measurement of clinically important parameters measured from the temporal distribution of the ECG constituent waves we shall review non-linear transformations of the ECG, the use of principal component analysis (linear and non-linear), ways to map the transformed data into *n*-dimensional spaces, and the use of neural networks (NN) based techniques for ECG pattern recognition. This paper gives a non dimensional accurate information and the Main disadvantage is It cannot be easily predict

III. OUR APPROACH

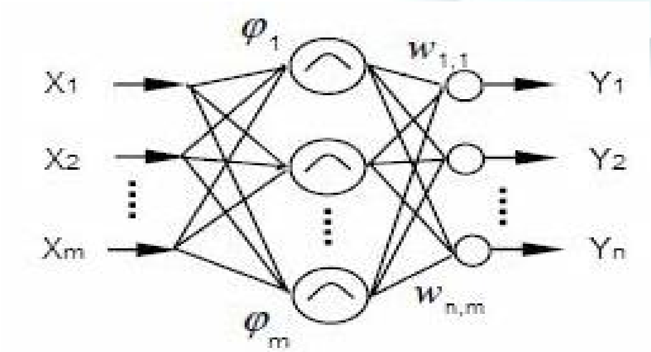
OBJECTIVE:

In this work, contributes the Big Data Reduction Using Distributed Radial Basis Function Neural Network RBFNN is proposed.

- To predict the kidney failures in Clinical Decision Support Systems based on the Distributed Radial Basis Function Neural Network (DRBFNN)

IV. BACKGROUND

Distributed Radial Basis Function Neural Network (DRBFNN) to predict the kidney failures . Algorithms mainly in two categories (1) descriptive or unsupervised learning and (2) predictive or supervised learning (i.e., classification, regression). However, they are lacking deeper insight into the suitability of the algorithms for handling the special characteristics of the sensor data in health monitoring systems. This model revealed to be accurate and suitable when applied on healthcare and wellbeing context. This model presented by taking large number of patient records with the parameters of age , sex, hb , esr , bp , br , albumim , urea, sodium bicarbonate, rt kidney,lt kidney



A radial basis function is a multidimensional function that describes the distance between a Given input vector and a pre-defined center vector. There are different types of radial basis

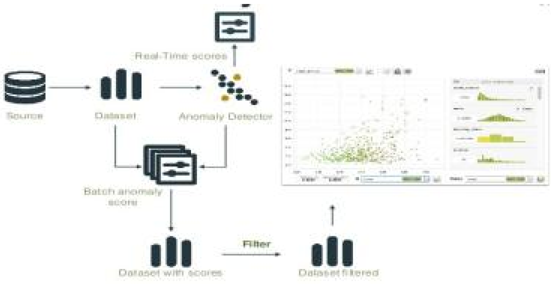
Function. A normalized Gaussian function usually used as the radial basis function, that is

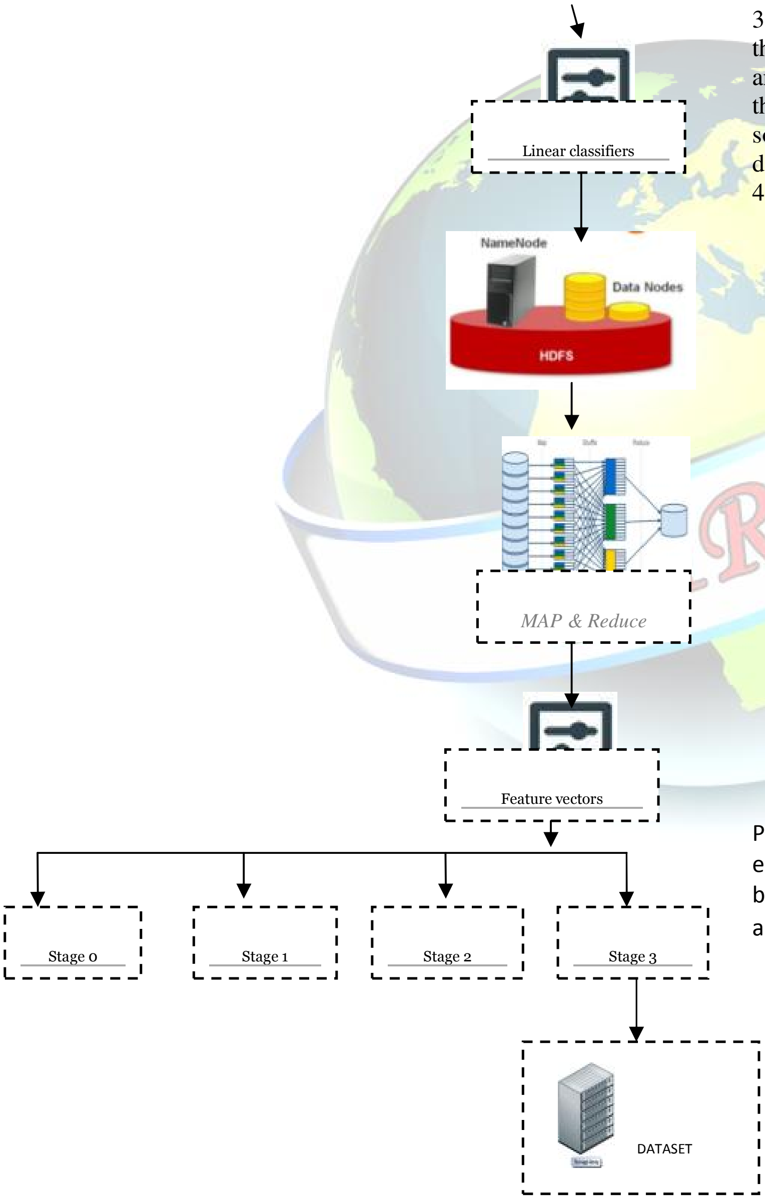
$$f(x)=\sum_{i=1}^{n} c_i\, G(||x-t||_{w})$$

and $G(||x-t||_w)$ - Radial basis function

A.Architecture diagram

The data from the source is uploaded in the dataset and then these datas are then analyzed for the Anamoly Detector and Batch Anamoly Score ,Anomaly Detection, which can help automate a number of predictive tasks for fraud detection, security, quality control, diagnoses and more. . and Real Time Scores **Batch Anamoly score** helps increase performance of other modeling tasks.Effective fraud detection must sift through huge volumes of data coming from many different sources, both inside and outside the company .**Real time scores** the company the option both to execute antifraud rules to identify suspicious links and to graphically visualize those connectionsThe datas form batch anamoly source are taken inside the Data inside with scores **Linear classifier**-a classifier that separates a set of objects into their respective groups with a line., **HDFS** in which the name nodes and data nodes are combined together and then mapper and reducer is done **Mapper**- the program needs to collect the output from the independent mappers and pass it to the reducers. This task is performed by an OutputCollector. **Reducer**-reduce gathers its output while all the tasks are processing. Reduce can't begin until all the mapping is done. The output of reduce is also a key and a value. OutputFormat takes the key-value pair and organizes the output for writing to HDFS. The last task is





B.MODULES

1.Anomaly detection- Anomaly detection is used to find the problem pattern in data .it does not give the expected behavior. Anomalies have two terms used in the context of anomaly detection . The main advantage of anomaly detection is that anomalies can translate the data according to the significant information in wide variety of applications

2.Prediction-After anomaly detection prediction is the next step. it is one of the processing component of insufficient resources.the input data rate exceeds the processing capacity of the component . the component contains many software bugs in it. Bottleneck shows all the information with the maximum cutoff

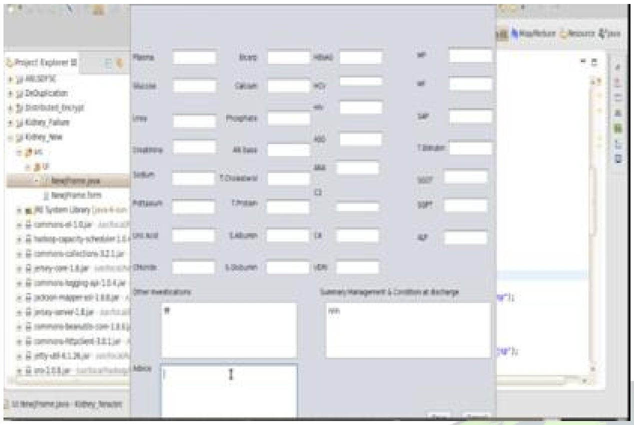
3.Diagnosis /Decision making-it is used to support different things of result and make decision .the decision must be strong and correct and produce results to discover the hidden things.in decision part small mistakes will make big problem so decions should be taken properly.the application can be done in any software

4.Other Big Data Mining Tasks

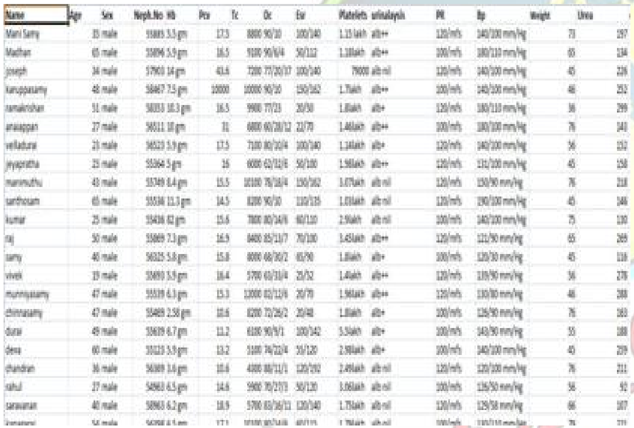
1. Preprocessing –it is the procedure to clean and transform the data before it is passed to the other models.It is used to remove the noise in the dataset and reduces the dimensions of space data cleaning is straightforward for applying the standard process of variance
2. Feature Extraction/Selection-It is used to measure the assets for arrangement of information. Examination for the substantial number of variables with large no of memory and calculation. The information contains elements that can either repetitive or immaterial without causing any loss of data

V. RESULTS AND DISCUSSION

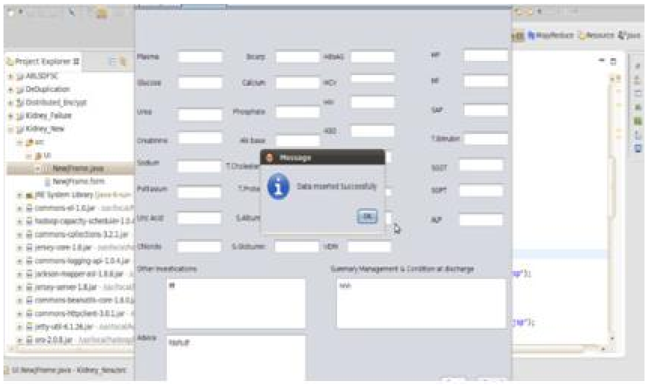
Patient test must be taken and then the the parameters are entered in text box i.eCretanine, Albumin,Sodium bicarbonate, Urea, Rt kidney, Lt kidney ,age,sex,potassium and platelets are uploaded.



Sampledataset



Name	Age	Sex	Neph.No	Hb	Pcv	Tc	Dc	Esr	Platelets	urinalaysis	PR	Bp	Weight	Urea
Mani Samy	35	male	55885	3.5 gm	17.5	8800	90/10	100/140	1.15 lakh	alb++	120/mts	140/100 mm/Hg	73	197
Madhan	65	male	55894	5.9 gm	16.5	9100	90/6/4	50/112	1.18lakh	alb++	100/mts	180/110 mm/Hg	60	134
joseph	34	male	57903	14 gm	43.6	7200	77/20/37	100/140	79000	alb nil	120/mts	140/100 mm/Hg	45	226
karuppasamy	48	male	58467	7.5 gm	10000	10000	90/10	150/162	1.7lakh	alb++	100/mts	140/100 mm/Hg	46	252
ramakrishan	51	male	58353	10.3 gm	16.5	9900	77/23	20/50	1.8lakh	alb+	120/mts	180/110 mm/Hg	36	299
arasappan	27	male	56511	10 gm	31	6800	60/28/12	22/70	1.46lakh	alb++	100/mts	180/100 mm/Hg	76	141
velladurai	23	male	56523	5.9 gm	17.5	7100	80/10/4	100/140	1.14lakh	alb+	120/mts	140/100 mm/Hg	56	152
jeyapratha	25	male	55364	5 gm	16	6000	63/32/6	50/100	1.98lakh	alb++	120/mts	131/100 mm/Hg	45	158
marimuthu	43	male	55749	8.4 gm	15.5	10100	78/18/4	150/162	3.07lakh	alb nil	120/mts	150/90 mm/Hg	76	218
santhosam	65	male	55536	11.3 gm	14.5	8200	90/10	110/135	1.03lakh	alb nil	120/mts	190/100 mm/Hg	45	146
kumar	25	male	55436	8.2 gm	15.6	7800	80/14/6	60/110	2.9lakh	alb nil	100/mts	140/100 mm/Hg	75	130
raj	30	male	55869	7.3 gm	16.9	8400	85/13/7	70/100	3.45lakh	alb++	120/mts	121/90 mm/Hg	65	269
samy	40	male	56325	5.8 gm	15.8	8000	68/30/2	65/90	1.8lakh	alb+	100/mts	120/30 mm/Hg	45	116
vivek	19	male	55693	5.9 gm	16.4	5700	63/33/4	25/52	1.4lakh	alb++	120/mts	139/90 mm/Hg	56	276
munniyasamy	47	male	55539	6.3 gm	15.3	12000	82/12/6	20/70	1.96lakh	alb++	120/mts	130/80 mm/Hg	46	288
chinnasamy	47	male	55489	2.58 gm	10.6	8200	72/26/2	20/48	1.8lakh	alb+	100/mts	126/90 mm/Hg	76	163
durai	49	male	55639	6.7 gm	11.2	6100	90/9/1	100/142	5.5lakh	alb+	100/mts	143/90 mm/Hg	55	188
deva	60	male	55123	5.9 gm	13.2	5100	74/22/4	55/120	2.98lakh	alb+	100/mts	140/100 mm/Hg	45	229
chandran	36	male	56389	3.6 gm	10.6	4000	88/11/1	120/192	2.49lakh	alb nil	120/mts	120/100 mm/Hg	76	211
rahul	27	male	54963	6.5 gm	14.6	5900	70/27/3	50/120	3.06lakh	alb nil	100/mts	126/50 mm/Hg	56	92
saravanan	40	male	58963	6.2 gm	18.9	5700	83/16/11	120/140	1.75lakh	alb nil	120/mts	129/58 mm/Hg	66	107
[illegible]	54	male	54398	4.5 gm	17.1	10100	80/14/8	60/115	1.78lakh	alb nil	100/mts	130/110 mm/Hg	75	221



Data inserted Successfully

Finally we get a dialog box with the message dataset uploaded in which the patient details are uploaded and then we further use CLASSIFIER and then HDFS to determine the stages of a patient affected by a kidney failure

VI. CONCLUSION

We conclude that it is useful for handling large volumes of clinical dataset for storing and retrieve it.The comparison of patient dataset results may promise to yield knowledge about patient records and to find the root cause of the kidney failures .In future it will be used to predict blood clots veins, brain tumour, weather forecasting, heart attack, Diabetics patient dataset. In Phase II, the complexity evolves due to data reduction can be overcome with the use of Map Reduce. So work of phase two is to implement DRBNN algorithm for big data reduction.Accurate results are produced by predictive modelling and it increases efficient handling of large volumes of patient dataset.

REFERENCES

[1] N. Pombo, P. Araujo, and J. Viana, "Knowledge discovery in clinical decision support systems for pain management: A systematic review," Artificial Intelligence in Medicine, vol. 60, no. 1, pp. 1-11, 2014.

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89-109, 2001.

[3] A. H. Gerardo and R. C. M. Antonio, "Cardiac Sudden Death Risk Detection Using Hybrid Neuronal-Fuzzy Networks," in Electrical and Electronics Engineering, 2006 3rd International Conference on, 2006, pp. 1-4.

[4]H. L. Lu, K. Ong, and P. Chia, "An automated ECG classification system based on a neuro-fuzzy system," in Computers in Cardiology 2000, 2000, pp. 387-390.

[5] D. J. Messadeg, C. Snani, and M. Bedda, "An approach for ECG classification using wavelets and Markov Model," in Information and Communication Technologies, 2006. ICTTA '06. 2nd, 2006, vol. 1, pp. 1910-1913.

[6] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," Anal. Chem., vol. 36, no. 8, pp. 1627-1639, Jul. 1964. "Evida, website da platforma, https://evida.pt/ (last access 5 May 2014)."

[7] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[8] T. Poggio and F. Girosi, "Networks for approximation and learning," Proceedings of the IEEE, vol. 78, no. 9, pp. 1481-1497, 1990. "ALLab-Assisted Living Computing and Telecommunications Laboratory, http://allab.it.ubi.pt/mediawiki/index.php/Main-Page (last access 5 May 2014)." , "Plux wireless biosignals, 'bioPlux', from www.plux.info/ www.plux.info/http://www.plux.info/ (last access 5 May 2014).".

[9] A. Godfrey, R. Conway, D. Meagher, and G. OLaighin, "Direct measurement of human movement by accelerometry," Medical Engineering & Physics, vol. 30, no. 10, pp. 1364-1386, Dec. 2008.

[10] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, "{ECG} pattern recognition and classification using non-linear transformations and neural networks: A review," International Journal of Medical Informatics, vol. 52, no. 3, pp. 191-208, 1998.