An Efficient Big Data Storage for Handling Kidney Failures Datasets using E-Health Platform Integration

G.Ben Sandra[1] V.L. Jyothi[2]
[1]M.E. Student [2]Head of Department
[1,2]Jeppiaar Engineering College, Chennai.

Abstract— Big data is an encompassing difficult or complex large datasets to process on traditional large scale data processing. The main confront of big data processing incorporates the extraction of significant data, from a high dimensionality of a wide assortment of medicinal information by empowering examination, disclosure and elucidation. These data are a useful tool for helping to understand disease and to formulate predictive models in different areas and support different tasks, such as triage, evaluation of treatment, and monitoring. In this work, based on a predictive model using the Distributed radial basis function neural network (DRBFNN) to aiming the estimation of kidney failures is presented. The proposed method exposed appropriateness to sustain patient & health care professionals (HCP) on clinical decisions and practices.
Key words: HCP, Radial basis function neural network, Hadoop

I. INTRODUCTION

Big data is an encompassing difficult or complex large datasets to process on traditional large scale data processing. The main confront of big data processing incorporates the extraction of significant data, from a high dimensionality of a wide assortment of medicinal information by empowering examination, disclosure and elucidation. These data are a useful tool for helping to understand disease and to formulate predictive models in different areas and support different tasks, such as triage, evaluation of treatment, and monitoring.

In this work, based on a predictive model using the Distributed radial basis function neural network (DRBFNN) to aiming the estimation of kidney failures is presented. The proposed method exposed it appropriateness to sustain patient & health care professionals (HCP) on clinical decisions and practices.

The branch of computer science which is more actively and efficiently involved in medical sciences is Artificial Intelligence. Various Clinical Decision Support Systems have been constructed by the aid of Artificial intelligence. These systems are now widely used in hospitals and clinics. They are proved to be very useful for patient as well as for medical experts in making the decisions.

Different methodologies are used for the development of those systems. The way of gathering the input data and to present output information's is different in different methodologies. Any computer program that helps experts in making clinical decision comes under the domain of clinical decision support system. An important characteristic of the Artificial Intelligence is that it can support the creation as well as utilization of the clinical knowledge. Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers. The quantity of data that is generated is very important in this context.

II. LITERATURE SURVEY

[1]Mokhairi Makhtar, Daniel C. Neagu, Mick Ridley has proposed Predictive Model Representation and Comparison: Towards Data and Predictive Models Governance.The procedures of creating prescient models include information planning, checking of data quality, lessening, displaying, expectation, and investigation of results. Creating superb prescient models is a time consuming activity because of the tuning process in finding optimum model parameters. Extraction of information mining models is an essential issue. Delivering the most helpful information mining models is inadequate without translating and handling learning from the models. This proposed work has an adaptable information and model representation in a more broad system towards information and model administration. Separating the model parameters from XML representation outlines that the models are profitable as far as understanding, and sensible for further utilization. PTML representation gives a helpful answer for separating information structure data mining.

[2]Andrea Ros`a, Lydia Y. Chen, Walter Binder proposed Predicting and Mitigating Jobs Failures in Big Data Cluster.Motivated by the significant amount of resource waste, in terms of computational time, CPU, RAM and DISK, caused by job failures at big-data clusters, the aim to capture failed jobs upon their arrival and minimize the resulting resource waste. To incorporate transient and complex system dynamics, we consider extensive static and system features that capture the disparities of jobs' multiple tasks and system load across priorities. The first explore four supervised classification techniques, namely LDA, ELDA, QDA, and LR, in a sliding window fashion, when developing an on-line prediction model for job failures. Based on the prediction results, the developed a delay-based mitigation policy that proactively terminates predicted-to-fail jobs after a certain grace period. The optimal choice of the classification technique, size of the sliding window, and length of grace period are determined during the training phase, so as to achieve low misclassification rates and mitigated false negative rates

[3]Nuno Pombo, Nuno Garcia, Virginie Felizardo, Kouamana Bousson proposed Big Data Reduction Using RBFNN: A Predictive Model for ECG Waveform for e-Health platform integration.This work highlighted the importance of methodologies for obtaining knowledge starting from the collected data and its capability to produce accurate and reliable outcomes for health care assistance professionals in the clinical decision making. In addition, the problematic of large volumes of a wide variety of clinical data was addressed. In line with this, are promising innovative technique which aiming to establish knowledge

refinement and discovery based on reduced datasets. Finally, a case study based on RBFNN combined with a filtering technique was presented. This model revealed to be accurate and suitable when applied on healthcare and wellbeing context. Additional studies should be addressed aiming to evaluate the combination of several parameters such as EMG, ECG and skin temperature, relating to the prediction of patients healthcare conditions.

[4]Seungwoo Jeon, Bonghee Hong and Byungsoo Kim proposed Big Data Processing for Prediction of Traffic Time based on Vertical Data Arrangement .This work is used for discovering new problems of predicting various traffic conditions according to time and location with historical traffic data for long-term prediction, and the problems indicate historical data aggregation and a variety of spatiotemporal traffic conditions. To solve the data aggregation issue, the proposed novel method called vertical data arrangement, which aggregates matching items of historical data into the same time slot. For a variety of spatio temporal traffic conditions & suggests constructing a spatiotemporal prediction map for each road and each day. By using the prediction map, the work can select suitable time-series forecasting methods for specific traffic conditions according to the location and time by analyzing the characteristics of historical data for each road and each day of the week. Moreover, both methods involve big data processing & constructed a big data processing Framework to handle the complete series of processes.

[5]Dr. Tariq Mahmud, Tasmiyah Iqbal, Farnaz Amin, Wajeeta Lohanna, Atika Mustafa proposed Mining Twitter Big Data to Predict 2013 Pakistan Election Winner.Twitter is a well-known micro-blogging website which allows Bmillions of users to interact over different types of communities, topics, and tweeting trends.The big data being generated on Twitter daily, and its significant impact.

[6] Niels Buus Lassen, Rene Madsen,Ravi Vatrapu proposed Predicting i-Phone Sales from i-Phone Tweets. Illustration from the hypothetical structure of AIDA and Hierarchy of Effects models in advertising combined with an assumption that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product, the proposed work how social media data from twitter can be used to predict the sales of i-Phones. The developed and evaluated work posses a linear regression model that transforms i-Phone tweets into a prediction of the quarterly i-Phone sales with an average error close to the established prediction models from investment banks. This strong correlation between i-Phone tweets and i-Phone sales becomes marginally stronger after incorporating sentiments of tweets.

[7]Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Brian Muckian implemented Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients.This work proposed the big data solution for predicting the 30-day risk of read mission.

[8]Jiang Zheng, Aldo Dagnino proposed An Initial Study of Predictive Machine Learning Analytics on Large Volumes of Historical Data for Power System Applications.Rising of industrial growth that combines computers, sensors, data repositories, high bandwidth networks, mobile devices, autonomous machines, and data analytics that drive industrial innovation and growth. More and more industrial data are being collected and stored by these industrial systems. For this reason, Industrial Analytics requires more powerful and intelligent machine learning tools, strategies, and environments to appropriately extract knowledge from the large volumes of industrial data to unleash its great potential value. This work started the research on predictive machine learning analytics for Big Data by conducting a comprehensive literature survey of machine learning libraries and tools for Big Data analytics, and initial studies on how to forecast substation faults and power loading. Final results indicated that it is feasible to forecast substations fault events and power load using Naïve Bayes algorithm in Map-Reduce paradigm or machine learning tools specific for Big Data .

[9]Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, Mihaela van der Schaar presented Mining the Situation: Spatiotemporal Traffic Prediction with Big Data.This work possesses a framework for online traffic prediction, which discovers online the contextual specialization of predictors to create a strong hybrid predictor from several weak predictors. The proposed structure matches the real-time traffic situation to the most effective predictor constructed using historical data, thereby self-adapting to the dynamically changing traffic situations. The systematically proved both short-term and long-term performance guarantees for our algorithm, which provide not only the assurance that our algorithm will converge over time to the optimal hybrid predictor for each possible traffic situation but also provide a bound for the speed of convergence to the optimal predictor. Final experiments on real-world dataset verified the efficiency of the proposed scheme and showed that it significantly outperforms existing online learning approaches for traffic prediction.

[10]Yang Xie, G¨unter Schreier, David C.W. Chang, Sandra Neubauer, Ying Liu, Stephen J.Redmond, Nigel H. Lovell proposed Predicting Days in Hospital Using Health Insurance Claims.Healthcare administrators worldwide are striving to lower the cost of care whilst improving the quality of care given. Hospitalization is the largest component of health expenditure. Therefore, earlier identification of those at higher risk of being hospitalized would help healthcare administrators and health insurers to develop better plans and strategies. In this paper, a method was developed, using large-scale health insurance claims data, to predict the number of hospitalization days in a population. This work utilized a regression decision tree algorithm, along With insurance claim data, to provide predictions of number of days in hospital.The proposed method performs well in the general population as well as in sub-populations. Results indicate that the proposed model significantly improves predictions over two established baseline methods

[11]Sudha Ram, Wenli Zhang, Max Williams, and Yolande Pengetnze proposed Predicting Asthma-Related Emergency Department Visits Using Big Data.Asthma is a common chronic inflammatory and its symptoms include wheezing, coughing, chest tightness, and shortness of breath. We introduce a novel method of using multiple data sources for predicting the number of asthma related

Emergency Department (ED) visits in a specific area. Our preliminary findings show that our model can predict the number of asthma ED visits based on near-real-time environmental and social media data with approximately 70% precision. The results can be helpful for public health surveillance, emergency department preparedness and targeted patient interventions.

[12]Marco Viceconti, Peter Hunter, and Rod Hose proposed Big Data, Big Knowledge: Big Data for Personalized Healthcare.A VPH full form is Virtual Physiological Human. The Virtual Physiological Human (VPH) is a methodological and technological framework that, once established, will enable collaborative investigation of the human body as a single complex system. The Virtual Physiological Human (VPH) is a framework which aims to be descriptive, integrative and predictive .we propose in this position paper that big data analytics can be successfully combined with VPH technologies to produce robust and effective in silicon medicine solutions.

[13]Yen Chen and Hue Yang proposed Heterogeneous Postsurgical Data Analytics for Predictive Modeling of Mortality Risks in Intensive Care Units.Intensive Care Units cater to patients with severe life-threatening illnesses and injuries, which require constant, close monitoring and support from specialist equipment, medications in order to ensure normal bodily functions. To cope with the challenges in ICU datasets, we developed the postsurgical decision support system with a series of analytical tools, including data categorization, data pre-processing, feature extraction, feature selection and predictive modeling.

[14]Bas Geerdink proposed A Reference Architecture for Big Data Solutions Introducing a model to perform predictive analytics using big data technology.With big data technology and predictive analytics techniques, organizations can now register, combine process and analyze data to answer questions that were unsolvable a few years ago. This paper introduces a solution reference that gives guidance to organizations that want to innovate using big data technology and predictive analytics techniques for improving their performance. The reference architecture is the result of an iteration of Hevner's framework for designing information systems artifacts.

[15]Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang has proposed Traffic Flow Prediction with Big Data: A Deep Learning Approach.An Accurate and timely traffic flow information is important for the successful deployment of intelligent transportation systems. A stacked auto encoder model is used to learn generic traffic flow features, and it is trained in a greedy layer wise fashion. Experiments demonstrate that the proposed method for traffic flow prediction has superior performance.

III. PROPOSED ISSUES

Anomalies are examples in information that don't fit in with a very much characterized idea of typical conduct. Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably. The importance of anomaly detection is due to the fact that anomalies in data translate to significant actionable information in a wide variety of application domains.

Anomaly detection discovers broad utilization in a wide mixed bag of uses, for example, misrepresentation location for charge cards, protection or human services, interruption discovery for digital security, deficiency identification in wellbeing discriminating frameworks, and military reconnaissance for foe exercises.

Anomaly happen because of various reasons, such as a processing component is given insufficient resources, the input data rate exceeds the processing capacity of the component, or the component contains software bugs (e.g., memory leak, buffer management error).Most widely recognized irregularity in information handling groups. An information stream handling application commonly comprises of an arrangement of preparing segments. Every part acknowledges info information from its upstream component(s) and produces yield information for its downstream component(s). A bottleneck shows up in the circulated application when the information line of a segment achieves its maximum cutoff.

The decision-making/ Diagnosis is supported by which inquire different things from the result, and then make various decisions .The decision must be strong and correct enough that efficiently produce results to discover hidden things and make decisions. The decision part is significant since any small error in decision-making can degrade the efficiency of the whole analysis. Finally, that any application can utilizes those decisions at real time to make their development. The applications can be any business software, general purpose community software, or other social networks that need those findings (i.e., decision-making/ Diagnosis).feature extraction and selection is used for classification of kidney failure dataset.

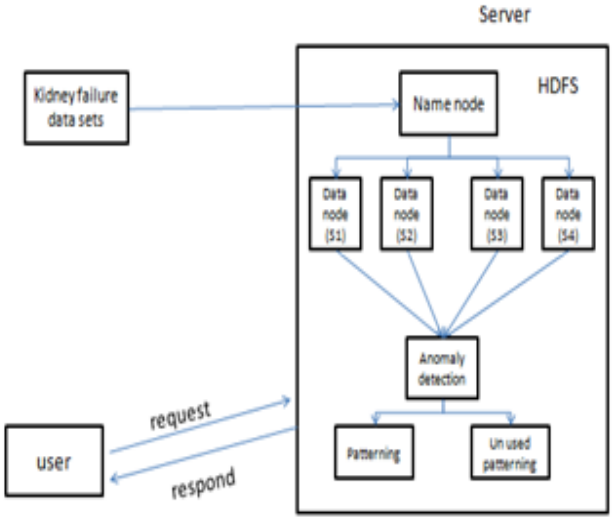

Fig. 1.1:

Preprocessing is a procedure to clean and transform the data before it is passed to other modeling procedure. Data cleaning involves removing the noise and outliers in the data set, while data transforming tries to reduce the irrelevant number of inputs, i.e., reducing dimensionality of the input space. As data cleaning is very straightforward of

applying standard process of zero mean and unit variance, the concentration is put on data transforming. The following subsections introduce the common data transformation method.

Feature extraction includes diminishing the measure of assets needed to depict a substantial arrangement of information. At the point when performing examination of complex information one of the significant issues originates from the quantity of variables included. Examination with a substantial number of variables by and large obliges a lot of memory and calculation force or an arrangement calculation which over fits the preparation test and sums up inadequately to new examples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

Feature selection method is that the information contains numerous elements that are either repetitive or immaterial, and can in this way be uprooted without causing much loss of data. Redundant or irrelevant features are two particular ideas, since one significant component may be excess in the vicinity of another pertinent element with which it is unequivocally associated.

IV. Related work

The RBF's are characterized by their localization and Gaussian activation function using supervised (gradient-based) procedures to obtain the expected result. In a supervised application, the network is provided with a set of data samples called training set for which the corresponding outputs are known. After training the network, to produce expected result, it is tested with another set of data samples called test set, to check whether the classifier has learnt to classify the given data effectively. Radial basis functions are embedded into a two-layer feed-forward neural network. Such a network is characterized by a set of inputs and a set of outputs. In between the inputs and outputs there is a layer of processing units called hidden units. Each of them implements a radial basis function.

In this study the inputs represent the feature entries and the output corresponds to a class. The hidden units correspond to subclasses in the neural network. The number of hidden units determines the classification accuracy of the network. The determination of number of hidden units is usually done using k-means clustering method. In the given architecture, there are p input vectors, k hidden layer units and q output units each unit corresponding to a class. The input of each RBF hidden unit is the linear combination of the input vector X = [x1, x2, xp] T and the scalar weights between an input layer and the hidden layer which is usually a unitary value. In the hidden layer, each hidden unit computes the activation of the weight vector cj associated with the jth hidden unit (represented by the jth column of a weight matrix C) and applies a radial symmetric output function f (typically a Gaussian function) to Xj.

In RBF networks, determination of the number of neurons in the hidden layer is very important because it affects the network complexity and generalizing capability of the network. If the number of the neurons in the hidden layer is insufficient, the RBF network cannot learn the data adequately; on the other hand, if the neuron number is too high, poor generalization or an over learning situation may occur. The position of the centres in the hidden layer also affects the network performance considerably. To determine the correct centre positions an unsupervised k-means clustering algorithm is used to partition the data into clusters. The k-means algorithm is one of the simplest learning algorithms to cluster n objects based on attributes into k partitions, k < n. To achieve good results, the RBF network requires a proper initialization of all weights cij which is done by the k-means clustering algorithm and of the width σj of the Gaussian function. After the initialization, the network is trained by a gradient descent training algorithm that adapts all weights cij, wjk and σj (the centre coordinates, heights and widths of Gaussian function) according to the error at the network outputs.

V. DRBFNN ALGOTHIMS

DRBFNN(Distributed RBFNN) algorithms mainly in two categories: descriptive or unsupervised learning (i.e., clustering, association, summarisation) and 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.

VI. Conclusion

This project we have proposed To handle large volumes of unstructured data for the storage management. To take good decision making for predicting acute kidney injury(AKI).To perform preprocessing and feature extractions for big data reduction using the Distributed Radial Basis Function Neural Network (DRBFNN) Algorithm.

To detect Anomaly Detection and Prediction for kidney failures. In this work, we study the big data solutions for predicting the Kidney failure and injury using DRBFNN Algorithm. To provide solutions for leverages big data infrastructure for both information extraction. An effectiveness for comprehensive set of experiment, considering the quality and scalability.

To leveraging big data infrastructure for our designed risk calculation tool, for designing more sophisticated predictive modeling and feature extraction techniques, and extending our proposed solutions to predict other clinical risks. In future it will be used to predict blood clots and brain tumour.

References

[1] Mokhairi Makhtar, Daniel C. Neagu, Mick Ridley School of Computing, Informatics and Media, University of Bradford, "Predicting Predictive Model Representation and Comparison: Towards Data and Predictive Models Governance"

[2] Andrea Ros`a and Lydia Y. Chen and Walter Binder " Predicting and Mitigating Jobs Failures in Big Data Clusters "2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

[3] Seungwoo Jeon , Bonghee Hong, and Byung soo Kim" Big Data Processing for Prediction of Traffic Time based on Vertical Data Arrangement" 2014 IEEE 6th International Conference on Cloud Computing Technology and Science

[4] Nuno Pombo, Nuno Garcia, Virginie Felizardo, Kouamana Bousson5" Big Data Reduction Using RBFNN: A Predictive Model for ECG Waveform for eHealth platform integration" IEEE HEALTHCOM 2014-The 2nd International Workshop on Service Science for e- Health (SSH 2014)
[5] Niels Buus Lassen, Rene Madsen, Ravi Vatrapu," Predicting iPhone Sales from iPhone Tweets" 2014 IEEE 18th International Enterprise Distributed Object Computing Conference.
[6] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy Brian Muckian, "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients," 2013 IEEE International Conference on Big Data.
[7] Jiang Zheng, Aldo Dagnino, "An Initial Study of Predictive Machine Learning Analytics on Large Volumes of Historical Data for Power System Applications," 2014 IEEE International Conference on Big Data.
[8] Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, Mihaela van der Schaar, "Mining the Situation: Spatiotemporal Traffic Prediction with Big Data," IEEE Journal of Selected Topics in Signal Processing.
[9] Yang Xie, G¨unter Schreier, David C.W. Chang, Sandra Neubauer, Ying Liu, Stephen J. Redmond, Nigel H. Lovell, "Predicting Days in Hospital Using Health Insurance Claims" IEEE Journal of Biomedical and Health Informatics.
[10] Sudha Ram, Wenli Zhang, Max Williams, and Yolande Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data" IEEE Journal of Biomedical and Health Informatics
[11] Marco Viceconti, Peter Hunter, and Rod Hose" Big data, big knowledge: big data forpersonalised healthcare" IEEE Journal of Biomedical and Health Informatics.
[12] Yun Chen and Hui Yang" Heterogeneous Postsurgical Data Analytics for Predictive Modelingof Mortality Risks in Intensive Care Unit".
[13] Bas Geerdink" A Reference Architecture for Big Data Solutions Introducing a model to perform predictive analytics using big data technology" The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013).
[14] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li and Fei-Yue Wang "Traffic Flow Prediction With Big Data:A Deep Learning Approach"IEEE Transactions on Intelligent Transportation Systems.
[15] Dr. Tariq Mahmood, Tasmiyah Iqbal, Farnaz Amin, Wajeeta Lohanna, Atika Mustafa "Mining Twitter Big Data to Predict 2013 Pakistan Election Winner".

