Predictive Analytics on Big Data - an Overview

Gayathri Nagarajan and Dhinesh Babu L.D
School of Information Technology and Engineering, Vellore Institute Of Technology, Vellore, India
gayunagarajan1083@gmail.com
E-mail: lddhineshbabu@gmail.com

Overview paper

Keywords: predictive analytics, big data, machine learning

Received: November 5, 2018

Big data generated in different domains and industries are voluminous and the velocity at which they are generated is pretty high. While research works carried out continuously to handle big data is at one end, processing it to develop the business insights is a hot topic to work on the other end. Though there are lot of technologies and tools developed to handle big data and extract insights from them, there are lot of challenges and open issues that are yet to be addressed. This paper presents an overview on predictive analytics with big data. The overview throws light on the core predictive models, challenges of these models on big data, research gaps in several domain sectors and using different techniques. This paper categorizes the major technical challenges of predictive analytics on big data under six headings. The paper concludes with the identification of open issues and future directions to work on for researchers in this field.

Povzetek: Pregledni članek opisuje prediktivno analitiko na velikih podatkih.

1 Introduction

Research focus on predictive analytics for big data has gained significance because of its scope in various domains and industries. It has stepped into every field including health care, telecommunication, education, marketing, business, etc. Predictive analytics is a common diction that often means predicting the outcome of a particular event. The main idea behind prediction is to take certain input data, apply statistical techniques and predict the outcome of an event. The terminology 'predictive analytics' is synonymous with other terminologies like 'machine learning', 'data mining', 'business intelligence' and recently the other terminology which is in common use today 'data science'. Though they seem to be synonymous there is a narrow line that distinguishes their context of use.

The technique of business understanding, data understanding, data integration, data preparation, building a model to extract hidden insights, evaluating the model and finally deploying the model is called 'Data mining'. The model may be predictive or may not be. [1]. In some cases it may be descriptive whereas 'predictive analytics' in most cases mean to predict the value of certain output variable from input variables. 'Machine learning' is basically a technique whereas 'predictive analytics' is an application of machine learning. 'Machine learning' is used to discover hidden patterns in data by using some of their techniques like classification, association or clustering in training the machine. 'Machine learning' is one disciplinary of 'data mining' which is multidisciplinary that includes other disciplines like statistics, BI tools, etc. 'Data science' can be considered as an application of statistical methods to business problems. Predictive analytics is more narrowly focused than data science. Data science uses data programming whereas predictive analytics uses modeling. Predictive analytics in most of the cases is probabilistic in nature whereas data science involves exploration of data. Data scientists require both domain knowledge and the knowledge in technology. Business intelligence provides standard business reports, ad hoc reports on past data based on OLAP and looks at the static part of the data. Predictive analytics requires statistical analysis, forecasting, and causal analysis, text mining and related techniques to meet the need of forward looking business [1].

In predictive analytics, data is collected from different input sources. A model is developed based on statistics. The model is used to predict the outcome after proper validation. With the advent of big data, predictive analytics on big data has become a significant area of interest. Though there are lot of tools and techniques available to handle predictive analytics on big data, there are yet challenges open for the researchers to work upon. Our paper aims to present an overview on predictive analytics in big data to aid the researchers understand the contemporary works done in this area thereby providing them research directions for future work. We focused on including the research works carried in different industries and using different techniques so that the researchers can focus more on their specific area of interest after a complete understanding of the works done in different fields and using different techniques. The mo-

tivation behind this work is the fact that many papers in this field are more focused on a particular domain or technique but there is a lack of papers that presents a broader overview of predictive analytics in big data to help the budding researchers identify research problems. Hence we focussed on a comprehensive overview on predictive analytics.

This paper is organized into 7 sections. Core predictive models with their strengths, weaknesses along with few solutions are discussed in Section 2, the challenges of core predictive models on big data is discussed in Section 3, scope of predictive analytics on big data generated across different domain sectors along with few research gaps is discussed in Section 4 and the comprehensive challenges for predictive analytics on big data and the techniques used to overcome them is discussed in Section 5, the future directions for research are summarized in Section 6 and Section 7 winds up with conclusion.

2 Core predictive models

The major processes of predictive analytics include descriptive analysis on data that constitutes around 50% of the work, data preparation(like detecting outliers) that constitutes around 40% of the work, data modeling that constitutes around 4% of the work and evaluating the model that constitutes around 6% of the work [98]. Only a fraction of raw data is considered for building the model which is assessed and tested [7]. The phases involved in building predictive models is shown in figure 1. Initially predictive analytics was carried out using many mathematical statistical approaches. Later data mining, machine learning began its era in predictive analytics since they proved to be effective. This section discusses few core predictive models to make the reader understand the concept of predictive analytics. Different models are used for different types of predictive tasks such as estimating an unknown value, classification(supervised learning to predict the class label of the instance), clustering(unsupervised learning to group similar patterns together) etc. The section is branched into three subsections - the predictive models based on mathematical(statistical) approaches, the models based on data mining approaches and the models based on machine learning approaches respectively. Yet, there is a very narrow line of separation among the subsections and they overlap in certain predictive tasks. Figure 2 shows the classification of core predictive models.

2.1 Predictive models based on mathematics

Mathematical techniques especially statistics is used for predictive tasks. Despite, data mining algorithms and machine learning algorithms also use math as their base for predictive tasks. Major core predictive models based on mathematics include Extrapolation, Regression, Bayesian statistics that are described in detail.

2.1.1 Extrapolation

Extrapolation is a method of extending the value of a variable beyond the original observation range. For example, the details of the road condition are known to a driver until a certain point and he is expected to predict the road condition beyond that point. A tangent line is drawn by relating the input variable to the output variable. The line is extended further to predict the output for different values of input. The line determines whether the extrapolation is linear, quadratic or cubic etc.

Strength and weakness :

Extrapolation suits well for such tasks where the target variable is in close relationship with the predictor variables. The results of extrapolation are also accurate in certain experiments where the relationships among the variables are simple [101].

The major problem with extrapolation is the interpretation of results. There are many studies where the study population differs widely from the target population.[100] is an example of such problem where the extrapolation of the experimental results on sheep cannot be justified for other target population. In such studies, the claims of the study results cannot be applied or justified to the target population [99]. Few solutions proposed to solve the interpretation problem of extrapolation is simple induction, randomized trails and expertise, Mechanistic reason etc. Population modeling is also proposed to solve the extrapolation problem [102]. But these solutions can help only to a certain extent. Secondly, it is hard to model the past with extrapolation. Sometimes several extrapolation methods are combined to model. Moreover, extrapolation cannot be used to model the tasks with non linear patterns [103].

2.1.2 Regression

Regression models are used for supervised learning problems in predictive analytics. It works by establishing a mathematical relation of the input variables with the output variable. There are different types of regression models like linear regression model, multi variate regression model, logistic regression model, time series regression model, survival analysis, etc. depending on the nature of the relationship discovered among the variables. Though the term is synonymous with extrapolation, there is a difference. Regression explains the variations in the dependent attribute with respect to the variations in the predictor attributes. Also, regression doesn‘t use the value of the input variables outside the range to establish the relationship with the dependent variable as in the case of extrapolation. There are different variants of regression depending on the nature of the variables as shown in table 1.

Strength and weakness:

Linear regression can suit well on tasks where the variables exhibit linear relationship [104]. For predictive tasks where associated probability is also important apart from predicting the value of a variable such as in [110], logistic regression is preferred. For predictive tasks where a

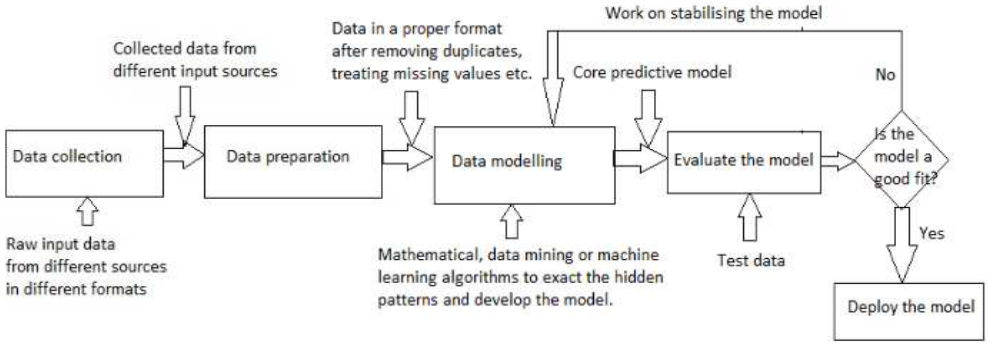

Figure 1: Phases involved in building a core predictive model

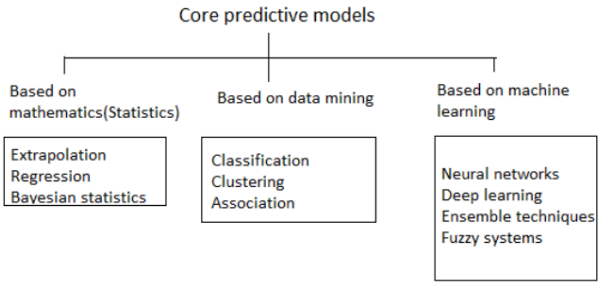

Figure 2: Classification of core predictive models

non parametric and non linear methodology has to be used, MARS regression can be considered as it does not assume any functional relationship between the dependent and independent variables [114],[117]. The biggest strength of MARS is that it is simple to understand and easy to interpret. Despite, it does not require any data preparation [116]. Moreover it is suitable for problems with higher input dimensions due to its 'divide and conquer' strategy of working principle [117]. To model complex processes, in which no theoretical models exist, LOESS regression is preferred as it does not require a function to fit a model to all data in the sample [121]. The major strength of LOESS regression is its flexibility. The flexibility is controlled by a smoothing parameter [119]. It is usually set between 0.25 and 0.5 [120]. For predictive tasks, where the number of predictors is more, regularization methods such as ridge regression is preferred as it avoids overfitting [122]. It also handles multicollinearity effectively [126].

The major weakness of linear regression is that it considers the mean of dependent variable and hence is sensitive to outliers. It is also more suited only to applications in which the predictor variables are independent [104]. For example, [105] states the reasons for moving to machine learning predictive models from simple regression models for predictive tasks in medicine. Though the regression models are simple and robust, they are limited to a small number of predictors that operate within a range. Outlier detection algorithms are proposed for linear time series regression models [106],[107]. Another problem with linear regression models is that it does not suit for predictive tasks when the predictor variables are collinear. [109] is an example that shows the adverse effect in interpretation of results when regression is performed without considering the multicollinearity problem. Techniques such as principal component analysis or stepwise regression can help to remove highly correlated predictors from the model [108]. The major weakness with logistic regression is that it is affected by omitted variables. Solutions such as replacing the latent continuous variable with an observed continuous one is proposed. Moreover,the odds ratio obtained from logistic regression cannot be interpreted easily as the heterogeneity of the model is not accounted [112]. In clinical predictive tasks, the odds ratios are estimated as risk ratios which is actually an overestimation. Alternatives such as Mantel–Haenszel risk ratio method, log–binomial regression, Poisson regression are used to give correct risk ratios [113]. Interpretation of results can be achieved to a certain extent by observing the probability changes [111]. Logistic regression is also not found to perform well with class imbalance problems and in such cases algorithmic approaches such as random forests is used [112]. The major weakness of MARS regression is overfitting and methods such as generalized cross validation is used [115],[118]. Pruning technique is also used to avoid overfitting problem to some extent by reducing the number of its basic functions thereby limiting the complexity of the model [117]. LOESS regression is less sensitive to outliers yet they are also overcome by extreme outliers. Another disadvantage with LOESS regression is that it requires densely sampled data to produce good results [121]. The major problem with ridge regression is the parameter settings. The hyperparameter has to be set [124]. Few methods are proposed in [123], [128] for parameter setting. Ridge regression also suffers from interpretability [124]. This happens because the unimportant predictors still exists in the model with the coefficients close to zero but not exactly zero [125]. LASSO regression is preferred in such predictive tasks to avoid the inter-

S.no	Type of regression	Explanation
1	Linear regression	The linear regression is given as $$Y = aX + b \quad (1)$$ where Y represents the predictor attribute, X represents the independent attribute, a is the slope and b, the intercept. The best fit line is obtained by minimizing the square of variations of each observed and the actual values with the line.
2	Logistic regression	Logistic regression is used for problems that are binary in nature and hence is mainly used for classification. This method aids to determine the likelihood of the occurence of a happening. It is denoted by $$logit(a) = ln(\frac{a}{1-a}) \quad (2)$$ where a is the likelihood of the occurence of the event.
3	MARS	Multivariate adaptive regression splines represented as MARS uses stepwise regression for model building. It is a non parametric method. The non linearities among the variables are identified and modelled with hinge functions. These functions create a hinge in the best fit line automatically according to the non linear relationship among the variables. The MARS equation is given as $$D = \beta_0 + \sum_{n=1}^{N} \beta_n h_n(I) \quad (3)$$ where D represents dependent attribute, I represents independent attribute, β_0 represents intercept variable, β_n represents slope of the hinge function $h_n(I)$.
4	LOESS regression	LOESS regression is a non parametric regression model. This method helps to fit regression line on subset of data rather than the data as a whole. It incorporates the concepts of regression model along with the nearest neighbor concepts.
5	Ridge regression	Ridge regression is another method commonly applied when the dataset experiences multicollinearity. They reduce the standard errors. A shrinkage parameter λ is added to the least squares term to minimize the variance.Ridge regression estimator is given as $$\beta_{ridge} = (I^T I + \lambda I_p)^{-1} I^T D \quad (4)$$ where I represents independent attribute matrix, D represents predicted attribute matrix, Ip represents identity matrix and λ represents shrinkage parameter.

Table 1: Different types of regression

pretability problem of ridge regression as it sets the coefficients of unimportant parameters to zero [127].

2.1.3 Bayesian statistics

Bayesian statistics predicts the likelihood of events occurring in the future. It works in the same way like normal probability but uses the input from experimentation and research to adjust the beliefs. For example, the probability of a six occurring in a die thrown six times is 1/6. But Bayesian statistics starts with the initial value of 16.6% and then adjusts the belief based on the experimentation. If the die is showing 6 more than the expected number of times during experimentation, the value of this belief is increased accordingly. Hence the likelihood of 6 turning in a die thrown 6 times will increase or decrease depending on the outcomes in the experimentation.

Strength and weakness:

Bayesian approaches yield accurate results in many predictive tasks as they consider both experimental data and theoretical predictions [129]. Bayesian approaches are best suited for tasks where there are uncertainties in models and parameters. They also find their use in predictive tasks where probabilities questions have to be answered such as

in stock market analysis [130].

The major weakness of bayesian approaches lies in determining the prior distributions. Bayesian approaches are also computationally expensive [130]. Use of frequencies instead of probabilities can help in improving bayesian reasoning [131].

2.2 Predictive models based on data mining

The process of extracting hidden patterns from the given input is called data mining. It is basically mining knowledge from the data. Three major approaches for data mining include Classification, Clustering and Association. Machine learning algorithms are widely used to execute the task.

2.2.1 Classification

The method of determining the class to which a particular data given as input belongs to is called classification. It is a supervised machine learning technique with labelled input data. Classification can be used in predictive analytics. There are lots of algorithms under classification technique but few basic algorithms are described in detail in this subsection.

1. Naive Bayes: Naive Bayes is a statistical modeling approach with two assumptions —all the attributes are equally important, all the attributes are statistically independent. It is a probabilistic approach which works on the following Bayes theorem.

$$P[\mathrm{M}/\mathrm{N}] = \frac{P[\mathrm{N}/\mathrm{M}]P[\mathrm{M}]}{P[\mathrm{N}]} \tag{5}$$

Strength and weakness:

The main strength of naive bayes is its simplicity. Inspite of the fact that its accuracy is less, it is found to perform better due to its simplicity in tasks such as document classification where merely classification is important. Naive bayes is computationally efficient since the contribution of each variable is equal and independent [134]. Moreeover only few parameters need to be calculated in naive bayes due to its conditional independence assumption. Hence it suits well for tasks where the training data is less [135].

The major weakness of naive bayes approach is its conditional independence assumption that often does not hold for real world applications [132]. This weakness is overcome to a certain extent by weighting attributes. Naive bayes with attribute weighting is found to perform better than random forest and logistic regression in [136],[137]. Accuracy and speed of classification is also less when compared with other classification approaches. Effective negation and feature selection techniques such as mutual information in combination with naive bayes is found to improve the accuracy and speed to a certain extent[133]. An another problem with naive bayes is that though they are good classifiers, they are not good estimators as discussed earlier. Hence it does not perform well in the tasks where probability value is important [138] and certain improvements to naive bayes is proposed to improve the probability estimation [139]. Moreover when the test data differs widely from the training data naive bayes fails to perform well unless smoothing techniques such as laplace estimation is used [138].

2. Decision trees: Decision tree is constructing a tree based structure for classification. Each node involves testing a particular attribute and the leaves are assigned classification labels based on the values at each node. Decision trees use divide and conquer approach and the attributes can also be selected with heuristic knowledge for the nodes of decision trees though few measurements like entropy and information gain are used in selecting the attributes. Decision trees can be converted to rules also. Many variations of decision trees have evolved and one of them is random forest which is commonly used bagging method in recent research problems. The leaf nodes of the decision trees are called decision nodes. Entropy is the amount of randomness in the data. Information gain is the information obtained that helps for accurate prediction. Entropy is given by

$$E(X) = -\sum_{i=1}^{n}(P_i log P_i) \tag{6}$$

where $\mathrm{P_i}$ is the probability of occurence of value i when there are n different values.

Information gain is a purity measure given by

$$IG(X, a) = E(X) - E(X|A) \tag{7}$$

The value represents the information gained by splitting the tree with that particular attribute. The attribute with less entropy or more information gain at every stage is considered the best split and the process is repeated until the tree converges. There are many decision tree algorithms but few variants are shown in table 2.

Strength and weakness:

The major advantage of decision tree over other classifiers is its interpretability. The tree like structure helps users to extract the knowledge easily [140]. Indeed, decision trees does not require the data to be normally distributed. The data can be continuous, discrete or a combination of both. Hence there is no need for much data preparation in decision tree model [142]. Moreover decision trees require only very few training iterations [143]. The random forest, an ensemble technique of decision tree is found to yield more accurate results. An another advantage of random forest is that it is non parametric in nature and helps in determining variable importance [141].

S.no	Type of decision tree	Description
1	ID3	Iterative dichotomiser is the basic non incremental algorithm used for construction of decision tree. It uses information gain measure to select the best split at each node. But the major disadvantage of ID3 is that it may overfit the training data and may not result in optimal solution.
2	C4.5	C4.5 is an improved version of ID3 algorithm. It solves the overfitting problem of ID3 by pruning the tree. C4.5 handles both continuous and discrete attributes and is capable of handling missing values too.
3	C5.0	C5.0 is an improved version of C4.5 in terms of performance and memory efficiency. It supports boosting technique to improve accuracy. C5.0 constructs smaller decision trees by removing unhelpful attributes without compromising on accuracy.
4	Decision stumps	It is a single level decision tree and finds its use along with machine learning techniques like bagging and boosting as weak learners.
5	CHAID	Chi square automatic interaction detector is used to find the relationship between categorical dependent variable and categorical independent variables. It uses chi square test of independence to test the independency between two categorical variables. CHAID can be extended for continuous variables too.

Table 2: Different types of decision trees

The major problem with decision trees is overfitting or underfitting [144]. Techniques such as pruning [146],[147] or feature selection methods are required to avoid overfitting problem and also to reduce the computational complexity of the model [147]. Decision trees does not suit well for imbalanced datasets though ensemble techniques can help [145]. Moreover, though random forest yields more accurate results, they are black box classifiers as the split rules are unknown [141].

3. KNN: Another classification technique that is of wide use is KNN yet with its own challenges. This technique works on the idea that the input data to be classified depends on the class of its neighbors. The value of 'K' determines the effectiveness of the algorithm. 'K' represents the number of neighbors to be considered. The input data is assigned to the class to which most of its neighbors belong to. A distance metric from the input data to the 'K ' neighbors is calculated. Euclidean distance is usually deployed to calculate the distance. Other distance metrics like mahanalobis and manhattan distance measures can also be used instead of euclidean. The accuracy of the algorithm lies in the choice of K. Lower value of K might result in overfitting and higher value for K might result in a more generalized model difficult to predict.

   Strength and weakness:

   The biggest advantage of KNN is its simplicity [150]. It does not require any prior knowledge about the distribution of data [149]. This non parametric nature of KNN makes it effective for real world problems [151].

   The major issue with KNN is the choice of parameter k and distance metric [150],[152],[153] and few works are proposed to determine the value of k [157],[158] etc. Computational complexity is another issue with KNN. Techniques such as clustering are used along with KNN [148] to reduce the computational complexity. KNN is also affected by irrelevant features [150] and is sensitive to outliers [152],[153]. The outlier problem can be avoided to a certain extent by choosing a reasonable value for k rather than a small k [154]. Few methods or improvements such as local mean based KNN [155] and distance weight KNN [156] are proposed to overcome the negative effect of outliers in KNN.

4. Support vector machines: This classification technique yields better accuracy in classification problems. SVM works on the basis of hyperplane. The idea behind this technique is to find the best hyperplane that can classify the two classes more accurately. This technique is best suited for both linear and non-linear separation problems. Non-linear problems can be handled using kernel functions that does data transformations to find the best hyperplane classifying the data more accurately. This algorithm works under the concept of margin. The distance between the hyperplane and the closest object in each class is calculated. The hyperplane with maximum margin with the class objects is the best classifier since it can predict the class more accurately. Each input data is assigned a point in n-dimensional space.

   Strength and weakness:

   The major strength of SVM is its robustness. It models non linear relationships very effectively and hence is proved to yield better results in terms of accuracy especially in non-linear problems [161][165]. SVM

is also known for its generalization capability and the generalization error is less in SVM [163] [164]. This advantage of SVM helps it to model complex problems even when the training data is less [166]. Moreover, there is no need for feature extraction process in SVM as the kernel function can be directly applied on the data [164]. SVM also avoids overfitting [165], [166].

The major weakness of SVM lies in its parameter settings. Proper setting of kernel parameters determine the accuracy of SVM. Certain optimization techniques such as PSO [159] and GA [160] are used to optimize the parameters of SVM. Methods such as double linear and grid search are also used to determine the parameter values of SVM [162].

2.2.2 Clustering

The technique of identifying similar patterns in the input data and grouping the input data with similar patterns together is called clustering. Clustering helps in predictive analytics. An example of clustering algorithm includes segmenting customers based on their buying behavior pattern thereby helping to predict the insights in the business and improve the sales accordingly. Clustering the scan images in health care helps to predict whether the person is affected by a specific disease or not. Though there are many clustering algorithms, the three basic algorithms include k-means, hierarchical and density clustering and a summary of the same is provided in the following subsection. A review of clustering with its scope, motivation, techniques and applications are explained in [2].

1. K-means clustering: K-means clustering chooses K random centroids and measures the distance of each input data point with the centroids. The most commonly used distance measurement metric is Euclidean distance. The input data points within the specific distance from the centroid are grouped together as a cluster and hence arrived at few clusters. The average of the distance of all the points from the centroid inside a cluster is calculated and the centroid is recalculated accordingly. The input data points belonging to the cluster changes again. This process continues until the centroids are fixed. Another variation of partition clustering is K-mediods where the centroid itself is an input data point. K-median and K-mode algorithms are also partition based clustering algorithms that uses median and mode instead of mean. There are several metrics to measure the performance of clustering. One among them is the distance metric. Single linkage is the nearest neighbor clustering where the minimum distance between the data points of two different clusters is calculated. Complete linkage is the farthest clustering where the maximum distance between the data points of two different clusters is calculated. Average linkage is also used in some scenarios.

   Strength and weakness:

   The major strength of k means clustering is it‘s simplicity and speed [168]. It can also work on datasets containing a variety of probability distribution [175].

   The major drawback with k means clustering is it‘s sensitivity to the initialization of cluster centers [167]. Hence determining the initial cluster centers is a major challenge though many methods based on statistics, information theory and goodness of fit are proposed [168]. Determining the number of centers is also a challenge and is addressed in few works [173]. An another drawback with k means clustering is it‘s computational complexity. As the distance of each data point has to be calculated with each cluster center for every iteration, the computational time is high. Solutions such as data structure that stores information at each iteration to avoid repeated computations [169] and Graphical processing units (GPUs) that parallelize the algorithm are proposed to reduce the computational complexity of k means clustering [172]. Moreover k means clustering is also sensitive to outliers and can end up in local optima. Few alternatives include fuzzy c means clustering and other soft clustering techniques that are proved to work well with noisy data [170]. k means clustering is also combined with optimization techniques such as PSO and ACO to avoid local optima and to arrive at better cluster partitions [171]. Few works are carried out to identify the better cluster partitions with minimum intracluster distances and maximum intercluster distances. Optimization function is derived that minimizes the intracluster distances and maximizes the intercluster distances. This function is optimized using optimization algorithms such as GA, PSO, WOA, ACO etc in few clustering works. Few other works include [269] that uses a set of objective functions and updates the algorithm accordingly to improve the intracluster compactness and intercluster separation, [270] that uses bisected clustering algorithm to measure the intracluster and intercluster similarity metrics etc. [174] proposes a method to overcome the drawback of noisy features in k means clustering.

2. Hierarchical based clustering: Hierarchical clustering works either in a divisive way (top-down) or agglomerative way (bottom-up). In the divisive clustering, large cluster is broken down into smaller pieces. In the agglomerative clustering, each observation is started as its own cluster and pair of clusters is merged together as they move up in the hierarchy. A dendogram is a pictorial representation for hierarchical based clustering. The height of the dendogram represents the distance between the clusters. Agglometric and divisive clustering algorithms are called AGNES and DIANA respectively. In DIANA clustering technique, all the input data points are considered as a single cluster and every iteration divides the cluster based

on heterogeneity. More heterogeneous data points breaks down into another cluster. In AGNES clustering technique, each input point is considered as a single cluster and homogeneous points are clustered together as a single cluster at each iteration.

Strength and weakness:

The major strength of hierarchical clustering includes it‘s scalability and capacity to work with datasets of arbitrary shapes [177]. It also determines the hierarchical relationships well. Moreover the number of clusters need not be specified in advance [178].

The major drawback with hierarchical clustering is its computational complexity [176],[177],[178] and few other methods are proposed to improve the efficiency of the same [176]. Parallel techniques are also used to improve the computational efficiency of hierarchical clustering [179].

3. Density based clustering: Density based clustering works by defining a cluster as the maximal set of density connected points. Three types of points are chosen core, border and outlier. Core is the part of the cluster that contains dense neighborhood. Border doesn‘t have many points but can be reached by the cluster. Outlier can‘t be reached by the cluster. Density based clustering picks up a random point and checks if it is the core point. If not, the point is marked as an outlier. Else, all the directly reachable nodes from the specific point are assigned to the cluster. It keeps finding the neighbors until it is unable to. There are certain kinds of problems where density based clustering provide accurate results than k-means clustering. Outlier detection is accurate in density based clustering.

Strength and weakness:

The major strength of density based clustering is that it can discover clusters of arbitrary shapes [184],[177]. It is also robust to outliers [184]. There are several density based algorithms such as DBSCAN, OPTICS, Mean-Shift etc [177].

The major drawback with density based clustering is the setting of parameters. Parameters such as neighbourhood size, radius etc. have to be set in density based clustering [182], [177]. Few algorithms are proposed to determine the parameters in density based clustering [183]. Moreover the density of the starting objects affect the behavior of the algorithm. The algorithm also finds its difficulty in identifying the adjacent clusters of different densities [182] ,[177]. Techniques such as space stratification is proposed to solve this problem [182]. An another drawback with density based clustering is its efficiency. Parallelization of the algorithm reduces the computational complexity to a certain extent. Techniques such as GPUs [180], mapreduce [181] are used to improve the scalability and efficiency of the algorithm.

2.2.3 Association

The method of identifying the relationship amid the items and deriving the rules based on the relationship is called association. Though association mining is not of much use in prediction, there are few scenarios where association rule is used. The rule has antecedent and consequent. Association rule mining is used mainly in business and marketing [185]. There are different algorithms used in association rule mining. Few include Apriori, Predictive Apriori, Tertius etc [186]. Optimization techniques such as PSO are also used with association rule mining to improve its efficiency [187]. [188] presents a survey on association rule mining algorithms.

2.3 Predictive models based on machine learning

Machine learning approaches are used for predictive tasks. It is the process of training the machine with a training input set, building a model and the evaluating it with the test data. The machine learns continuously from the errors until the model gets stabilized. Supervised learning works with labeled input data whereas unsupervised learning works with unlabeled input data. Machine learning uses soft computing techniques like neural networks for training.

2.3.1 Neural networks

Neural network is a commonly used soft computing technique for predictive analytics. Neural networks are used to classify complex patterns that are difficult to classify using SVMs or other techniques. There are different types of neural networks that can be trained using supervised, unsupervised and reinforcement learning. There are also different learning algorithms for training neural networks.

Neural networks machine learning algorithm can be used to train a network with a group of training data and then test it with a group of test data thereby measuring the accuracy of prediction. Learning continues until the network becomes stable and able to classify the data accurately. Cross validation is one among the widely used technique for evaluating the model. Backpropagation algorithm is the most commonly used training algorithm in neural networks. Weights are assigned at each layer input, hidden and the output. Weightage is given to each attribute based on the impact of it in predicting the output variable. Different types of functions like sigmoidal function and sign function are used to compute the output variable. These functions are called threshold functions and the output variable is predicted based upon these functions. Threshold functions are also called activation function or transfer function. The choice on number of input nodes, hidden layers, weightage, threshold functions, algorithm for learning are all based on the application and data for which predictive analytics has to be applied. Now, deep learning techniques are used to improve accuracy.

Strength and weakness:

The major strength of Artificial Neural Networks(ANN) lies in it‘s ability to work with large amounts of data and yield good results. They have good generalization and learning ability and are universal approximators [191]. ANN has good fault tolerant, self learning, adaptation and organization ability [192]. An another advantage with ANN is that they are good for high dimensional datasets as each variable do not have major impact on the class variable but as a group they are good at classification. Moreover a complex ANN relives user from determining the interactional and functional forms in prior and is able to smooth any polynomial function [193]. There are different types of neural networks such as as feedforward network, radial basis function network(RBFN), auto encoder, Boltzmann machine, extreme learning machines, deep belief network, deep convolutional network etc [189] each with its own strengths and weaknesses. For example, RBFN are easy to design, have good generalization ability and are tolerant to noise. These networks find their use in designing flexible structures and systems [190].

The major weakness with ANN is that they can‘t be applied blindly to all kinds of data. Hence they are used by combining with other models as hybrid prediction models in most of the prediction problems. For example, in time series problems, both linear and non linear relationships exist and ANN is combined with ARIMA modelling in such problems [190]. An another disadvantage lies in the fact that there are no proper rules to determine the number of hidden nodes in neural networks. Moreover they can also easily end up in local optima and are tend to overfit [193]. Optimization algorithms such as GA [194], Gravitational search algorithm with PSO [196] are used to avoid the local optima problem in ANN. Algorithms such as Fruitfly algorithm also find their use in determining the parameters for ANN [195]. Overfitting problems is addressed by techniques such as dropout mechanisms [197], bayesian regularization [198] etc.

2.3.2 Deep learning

Deep learning is the most commonly used technique in use today for classification, clustering, prediction and other purposes. While learning in machine learning proceeds in a broader way, deep learning works in a narrow way. It works by breaking down the complex patterns into simple smaller patterns. Learning happens in parallel in the smaller patterns and finally the sub solutions are combined together to generate the final solution. This improves the accuracy of the network. Deep nets also help in avoiding the vanishing gradient problem. Most of the deep learning problems use Rectified Linear units function(ReLUs) instead of sigmoidal and tanh activation functions that causes vanishing gradient problem. The use of ReLUs help overcome the vanishing gradient problem by avoiding zero value for the derivative and maintaining a constant value instead [271]. Moreover the use of deep learning networks such as Long Short Term Memory Networks(LSTM) avoid vanishing gradient problem by maintaining a constant value for the recursive derivative using cell states [272]. Deep nets use GPUs that help them get trained faster. When the input pattern is quite complex, normal neural networks might not be effective because the number of layers required might grow exponentially. Deep nets work effectively in solving such complex pattern by breaking them down into simpler patterns and reconstructing the complex solution from the simpler solutions. GPUs are known to be a better alternative to CPUs for big data analytics owing to it‘s lower energy consumption and parallel processing. Hence GPUs are found to be scalable in deep learning as the training and learning of the deep nets are made faster with parallel processing of many computational operations such as matrix multiplications [273]. There are different kinds of deep nets used for different purposes [5]. Table 3 shows the different types of deep nets and their usage.

Strength and weakness:

An overview of deep learning in neural networks has been discussed in [199]. The major strength of deep learning is it‘s ability to model non linear relationships well. Deep learning also suits well for massive data and has better representation ability than normal networks [200]. Moreover deep learning does automatic feature extraction and selection [201].

The major weakness of deep learning is that it is a black box technique. There is no theoretical understanding behind the model. Certain techniques such as information plane visualization are proposed to understand DNN by using the mutual information exchanged between layers in DNN [202]. Moreover deep learning works well only with massive data and their architectures are more specialized to a particular domain. They also consume high power [203].

2.3.3 Fuzzy rule based prediction

Fuzzy logic is a concept of soft computing technique more suited for prediction problems with uncertainty and imprecision. Fuzzy sets have membership functions associated with each input data set. The membership value of a particular input data represents the level of belonging of the particular input data to the particular set. Rules are derived and learning is based on the rules. Finally the rule based approach is used to classify or predict the output variable. Fuzzy systems are widely used for prediction purposes. Fuzzy systems can be used as stand-alone or can also be combined with other machine learning algorithms for predictive tasks. The simple fuzzy based classifier is If-THEN classifier and it can be made more meaningful with the use of linguistic labels [61],[204]. Fuzzy systems are also combined with neural networks as neuro-fuzzy classifier [261],[209] and is used for prediction purposes. Fuzzy systems are also combined with KNN for prediction purposes. Fuzzy c means clustering is found to perform well than hard clustering especially in applications such as bioinformatics where genes are associated with many clus-

S.no	Type of deep net	Description	Usage
1	Restricted Boltzmann machine	Two layered network with visible and hidden layer. Layers not connected among themselves. In the forward pass, RBM takes the input and encodes as numbers. The backward pass does the reverse. Data need not be labelled.	Recognize inherent patterns. Works well with real time data like photos, videos, voice etc. Used for clustering.
2	Deep belief nets	Stack of RBMs arranged together. Output of hidden layer of the previous networks like RBN is given as input to visible layer of next RBN.	Used for recognizing complex patterns. Used more commonly in facial recognition.
3	Convolution nets	Made up of three layers, convolution, RELU and pooling each having its own function.	Used to identify the internal patterns within an image.
4	Recurrent networks	A network with built in feedback loop. Uses techniques like Gating to overcome vanishing gradient problem.	Used when the patterns in the input data changes over time. For image captioning, document classification, classify videos frame by frame, natural language processing etc. LSTM is a recurrent network architecture that is used for deep learning. The application of LSTM includes time series data predictions, classification, processing, hand writing recognition, speech recognition etc. It is known for reducing the exploding and vanishing gradient problems.
5	Autoencoders	Encode the input data and reconstruct it to back. Works with unlabeled data.	Finds its use in dimensionality reduction. Used for text analytics.
6	Recursive Neural Tensor nodes	Works with the concept of roots and leaves. Data moves in the network in a recursive way.	Used to discover hierarchical structure of a set of data. Used in sentimental analysis. Used in natural language processing.
7	Generative adversial networks	A network that can produce or generate new data with the same statistics as the training data[274].	Used in fashion designing, improving satellite images, etc.

Table 3: Different types of deepnets and their usage

ters [170], [270].

Strength and weakness:

The major strength of fuzzy systems is it‘s interpretability. The fuzzy models are easy to interpret if designed carefully [205] especially with the use of linguistic labels [206]. Fuzzy rules also help to model the real world processes easily [207]. Fuzzy systems are known well for handling uncertainty [208].

The major weakness of fuzzy systems include it‘s poor generalization capability as it is rule based. Fuzzy systems are not robust as any change should be incorporated into the rule base. To overcome this disadvantage, fuzzy systems are often combined with ANN and hybrid systems are developed for prediction [208]. An another disadvantage of using fuzzy systems is that the knowledge about the problem should be known in advance. The use of hybrid systems can help overcome this disadvantage as the knowledge is extracted from neural networks in such systems [209]. Approaches such as genetic programming is also used to generate rules for fuzzy systems [210].

2.3.4 Ensemble algorithms

Ensemble methods are combination of more than one technique to achieve more accuracy in prediction than achieved by an individual model. Few ensemble techniques are shown in table 4. Each ensemble technique has its own strength and weakness. For example, bagging is stable against noise but needs comparable classifiers whereas boosting is unstable against noise but its classification performance is better than bagging [211]. Also, bagging is found to perform better than boosting for class imbalance problems especially in noisy environment [212]. Stacking has it‘s own weakness with respect to computational time. It is computationally expensive [213]. Another problem with stacking lies in the selection of base level classifiers as techniques such as exhaustive search consumes more time when search space is large. Yet, unlike bagging and boosting that uses the same algorithm, stacking uses a different algorithm and hence heterogeneous in nature [215]. The choice of the ensemble technique depends largely on the problem at hand. Bagging is good to deal with problems

S.no	Ensemble	Description
1	Bagging	Bagging or bootstrap aggregation is the method of decreasing the variance without any change in the bias. It is mainly used for regression and classification techniques. Each model is built separately and the net output is derived by bringing together the results from the individual models by joining, aggregation and other methods.
2	Boosting	Boosting is a parallel ensemble method to reduce bias without any changes in variance. Boosting converts weak learning algorithms to strong learning algorithms using certain methods like weighed average. There are many variations of boosting algorithms like adaboost, gradient boosting etc. The misclassified instances are assigned more weight in the successive iterations.
3	Stacking	Stacking is the technique in which the output of the previous level classifier is used as training data for the next classifier to approximate the target function. It minimizes variance and methods like logistic regression is used to combine the individual models.

Table 4: Ensemble techniques

where a single model is likely to overfit whereas boosting is good for problems where a model yields poor performance. Moreover bagging can be done in parallel as each model is independent whereas every model in boosting depends on the previous model [214]. There are different kinds of boosting techniques, the major include adaboost and gradient boost. Adaboosting improves performance by assigning high weight to the wrongly classified data points whereas gradient boosting improves performance by using gradients in the loss function [216]. Indeed, gradient boosting converges in a limit whereas adaboost is computationally efficient than gradient boosting [217].

3 Challenges of core predictive models on big data

'Big data' represents data sets that are in petabytes, zettabytes and Exabyte. The sources of big data include satellites that generate enormous information every second from space, mobile communications generating voluminous data, social media like Facebook, Twitter with blogs, posts etc. Traditional relational databases, data warehouses and many visualization tools and analytical tools are developed for structured data. Because of the heterogeneous nature of big data and enormous amount of data generated including real time data, there is a need to enrich traditional analytical methods to support the analytical functionalities for big data. Alternatively, new tools and techniques are developed to work on big data in combination with the traditional analytical techniques. Big data development includes the development in all the areas of handling big data including data storage, pre-processing, data visualization, data analytics, online transaction processing, online analytical processing, online real time processing, use of business intelligence tools for predicting insights from big data etc [3]. The main characteristics of big data include volume, velocity, variety, veracity and value.[4]. A single machine cannot store big data because of its volume. The basic concept behind big data storage is to have many nodes (computers) and store the chunks of big data in them. The nodes are arranged in racks and communicate with each other and the centralized node that controls them. Clouds are also used for big data storage but it has its own challenge of privacy and security. Getting into the depth of storage technology is outside the purview of this article and hence we are leaving the discussion about big data storage at this stage. As our paper mainly aims to discuss the overview of predictive analytics with big data, this section addresses the challenges encountered by the core predictive models discussed in previous section on big data.

Extrapolation

Extrapolation will be precise only when the knowledge about the underlying covariate information [220] and the actual system is clear [219],[221] which is difficult to determine in big datasets. With big data such as spatial data, existing extrapolation approaches fail due to it's time and space constraints. Hence new technological innovative approaches are required to model such big datasets and understand them [219]. Extrapolation with kernel methods like gaussian are proved to be good due to their flexibility in choosing the kernel function. Yet, when it comes to development of gaussian models for multidimensional big data, it suffers from computational constraints. Techniques such as recovering out of class kernels are used to overcome the computational constraint to a certain extent [218]. An another problem with the machine learning models including deep learning is that they merely fit the data and may perform well for training dataset and even testing dataset but fails in extrapolation [221],[222]. This happens as they do not have proper structural explanations for the correlations they identify [222]. [220],[221] recommends the construction of hybrid systems comprising the science based model or physical models along with the predictive models to improve the accuracy of extrapolation. But the problem in developing hybrid systems lies in the fact that it requires domain knowledge.

Regression

Regression is easy and can be understood well when the data is small and can be loaded into memory com-

pletely. But big datasets can not be loaded into memory completely. Few parallel techniques and solutions are proposed for regression yet they end up in local optima or in accessing the data again and again for updates. The other problem in using parallel techniques is the computational resources incurred [223]. The area in the improvement of computational resources is still lacking when compared with the amount of big data generated [224]. Regression approaches such as kringing is computationally complex especially with big data. Sampling techniques such as leveraging [224] and subdata selection[226] are proposed to reduce the computational complexity. But as discussed in the earlier sections, the inferences on the samples cannot be justified completely for the whole population as such. Regression is also performed locally by dividing the big dataset into few smaller datasets and then combining the submodels to construct the final model [225]. The challenges with these solutions lies in the choice of appropriate method for division, aggregation etc.

Decision trees

Big data streams are more prone to noise and decision trees are more sensitive to noisy data [227]. The time taken to build the decision tree is computationally expensive with big data [228]. Preprocessing and sampling the big data in full batches before the construction of decision tree adds to the computational cost [227]. External storage is required to construct decision tree for big data as the complete dataset cannot be loaded into memory. Hence tradition decision tree design does not suit for the big data. Solutions such as incrementally optimized decision tree algorithm [227] is proposed where decision tree is built incrementally. Parallel techniques [229] are proposed in big data platforms such as spark where the decision tree algorithm is executed in parallel. Decision tree algorithm is also converted into mapreduce procedures in [228] to reduce the computational time. The computational time of gradient boosted trees is decreased in [230] by eliminating few instances in calculation of information gain and bundling certain features together. Yet, these solutions come at the cost of choosing the right technique to break the algorithm for parallel execution, bundling the features etc.

K Nearest Neighbor

The major problem of KNN with big data is it‘s computational time as the distance has to be calculated among each instances [231]. This in turn incurs memory requirement for storage [232]. k means clustering is used to cluster the big dataset and KNN algorithm is used on each subset to reduce the computational time [231]. But this solution comes with the general limitations of k means clustering. Memory requirement is handled to a certain extent by big data platforms such as spark so that in time memory computation is used effectively [232]. Map reduce approaches [233] are also used to reduce the computational time. Parallelization of KNN algorithm is also proposed [234]. Yet, all these big data platform solutions come with their own concerns on the nature of partitioning as the accuracy can not be compromised for efficiency [235].

Naive Bayes

Naive Bayes requires the probability to be calculated for all the attributes. With big datasets, the number of attributes is more and hence the time complexity to calculate the probability for all the attributes is high [236]. Another problem with naive bayes is the underflow and overfitting problems [237]. The underflow problem is usually handled by calculating the sum of log of probabilities rather than multiplying the probabilities whereas overfitting problem is handled using techniques like laplace estimate, M-estimate etc. But with high dimensional big datasets like genomic datasets, these solutions are not efficient [237]. Naive bayes deals only with discrete data. Hence discretization methods are used before applying naive bayes algorithm. In case of big data, existing traditional discretization methods are not efficient and may lead to loss of information [238]. Parallel implementations are proposed for naive bayes algorithms yet they come at the cost of hardware requirements [236]. [237] proposes a solution to solve the underflow and overfitting problems in big data. The method uses a robust function that works based on average of condition probabilities of all attributes and calculation of dependency of the attributes on the class attribute. Parallel versions of existing discretization methods are also proposed to address the challenge of big data [239]. Yet, more research is required in these open issues.

Support vector machines

SVM is known for it‘s accurate results yet the computational complexity of SVM is quite high on big datasets [240],[241]. Inspite of this computational complexity, SVM uses certain optimization techniques like grid search method for parameter tuning. These optimization techniques are not suited for big datasets [244]. Though certain parameter optimization techniques such as stepwise optimization is proposed in [244] for big datasets, more research is needed in this area. Solutions such as implementing SVM on a quantum computer [240] to reduce it‘s time complexity is proposed. Again, they come at the cost of hardware. Parallel implementation of SVM using mapreduce technique is proposed [241] yet they may end up in local support vectors that may be far away from the global support vectors. [242] proposes a distributed version of SVM where global support vectors are achieved by retaining the first and second order statistics of the big dataset. Though there are many parallel versions of SVM, only a very few parallel tools are available in open source for parallel SVM. Moreover, these tools also require proper tuning [243].

K means clustering

The major problem of k means clustering with big data is it‘s computational complexity as the distance calculation and convergence rate incurs more time with increased number of observations and features. But it can be easily parallelized using big data platforms [245]. Though parallelization is easy with k means clustering using techniques like mapreduce, the I/O and communication cost increases due to repeated reading added with the iteration dependence

property of k means [246]. Methods like subspace clustering and sampling are used to reduce the iteration dependency property of k means [246]. Yet, the choice of correct sampling method and partitioning technique in case of subspace clustering adds to the big data challenges. Indeed, the size of the sample data is more than half the original data in most of the methods and hence the computational complexity still persists [248]. Optimization of initial clusters using techniques like choosing the data points in high density space [247] are proposed. Though they can avoid outliers, they still suffer from the same computational complexity owing to the distance calculation. Dimensionality reduction techniques are also proposed but they come with their own drawback that the cluster in projected space may not comply with the clusters in actual space [248]. Hybrid methods are proposed by combining projection techniques with sampling and few other techniques like visual assessment of cluster tendency. But the research on hybrid techniques are still in it‘s initial state [248].

Hierarchical clustering

Hierarchical clustering also suffers from the drawback of computational complexity and in fact it incurs more time than k means clustering when the size of the dataset is large [252]. Techniques such as building clusters using centroids [250] and usage of cluster seeding [252] are proposed to reduce the computational complexity of hierarchical clustering. Partitioning the sequence space into subspaces using partition tree is also proposed [251] and the clusters are refined in the subspaces. Fast methods to compute the closest pair is also proposed to reduce the computational cost. Yet, these methods are very specific to the particular problem. Moreover, the partitioning techniques and the cluster seeding techniques should be chosen wisely. Visual assessment of tendency are also used to return single linkage partitions on big datasets [249] yet the study of tendency curves have to be clear.

Density based clustering

Density based algorithms are better compared to partitioning algorithms on big data and data streams because it can handle datasets of arbitrary shapes. It is also not required to specify the number of clusters and it can handle noise effectively. But, with the high speed of evolving data streams and high dimensional big data, density based clustering is finding many challenges. Though few methods are found to perform better, they still suffer from open challenges such as too many parameters to set, memory constraints, handling different kinds of data such as categorical , continuous etc [254]. Big data platforms such as hadoop is used for parallelization in density based clustering. Yet, there is a need to choose the shuffling mechanism, partitioning technique and work load balancing efficiently [253]. Moreover, density based clustering algorithms such as OPTICS cannot be parallelized as such and either improvements or new algorithms have to be proposed to handle large datasets [255]. Few enhancements are carried out in OPTICS and other density based clustering algorithms to support parallelism. Yet, they are very specific to the problems they address. For example, [256] uses spatio temporal distance and temporal indexing for parallelization which is more specific to the spatio temporal data and [257] proposes a method that is specific to the electricity data.

Neural networks

The major challenge that neural network faces with big data is the time taken for training phase as large data sets require more training iterations or epochs [260],[261]. As a result, the computing power required becomes high [258] and in turn the energy consumption[260]. Though techniques like mapreduce on hadoop platforms are used [259], the mapreduce design has to be an optimized and efficient one. Hardware solutions such as GPU and memrister are proposed [258], yet they suffer from the major drawback, the cost factor. Optimization algorithms are proposed [259] to optimize the parameters of neural networks. Yet, there is a common perspective that using optimization algorithms increases the computational time due to it‘s convergence property though proper optimization decreases the training time of neural networks inspite of improving the accuracy. Very few researches are carried out in this area for big data with neural networks. The other problems with neural networks on big data include the increase in number of parameters, lack of proper theoretical guideline in the structure of neural networks, insufficient knowledge as only abstraction is extracted, inherent problem in learning etc[261].

Fuzzy systems

Fuzzy systems are of great use in big data due to it‘s ability to handle uncertainty. To cope up with the big data requirements, fuzzy systems are designed that distributes the computing operations using mapreduce technique [61]. But the major problem with mapreduce is the overhead taken to reload the job everytime during it‘s execution. Moreover when there are more number of maps, the imbalance problem has to be dealt with carefully. Spark which has in memory operations and resilent distributed databases is more efficient than mapreduce but unfortunately there are no big works that integrate fuzzy systems with spark [263]. Fuzzy systems are also found to be more scalable as they represent the data as information granules [262]. Yet, good granular techniques in combination with fuzzy classification systems exclusively for big data is required [262]. The fuzzy techniques designed for big data should be tested for real world problems and more general fuzzy techniques need to be developed rather than the techniques designed to address specific problems [262].

Deep learning

Deep learning is used for big data due to it‘s accuracy and automatic learning of hierarchical data representations [264]. Yet, the major problem with deep learning is the requirement of high performance systems. There are other areas that need to be explored in deep learning for big data problems. These include transfer learning with deep architectures, deep online learning that is still in the initial stage of research [264], incremental learning with deep architectures [265], working with temporal data [266] etc. Indeed, constant memory consumption due to the fact that

deep learning is usually performed on very big data that involves millions of parameters and many CPUs is another problem. Hence deep learning requires the use of excessive computational resources. Moreover, deep learning also faces the challenge of determining the number of optimal parameters, learning good semantic data representations as they are known for representing only the abstract detail etc [265]. Lot of research works is required to address the interpretablity problem[266].

Ensemble algorithms

The major problem with ensemble algorithms for big data is it's computational time [268]. As ensemble techniques require the use of different classifiers, the computational time they require is generally high and this increases when the data is big. Ensemble techniques are known for their diversity as they use different kinds of classifiers and aggregate the results. Though many ensemble techniques are developed for static data, there are no big research works carried out in studying the diversity of ensemble techniques on online streaming data. Since the different classifiers used on streaming data already differs in the data they use, a proper study of the advantages of ensemble techniques on streaming data is required. Proper pruning techniques is also an area to be explored [267]. There are few works where ensemble techniques for big data is parallelized with mapreduce [268], yet they are not tried on platforms such as spark that are proved to be more efficient than mapreduce.

4 Predictive analytics on big data across different domains

4.1 Healthcare

Big data is generated by lot of industries and health care is one among them. Huge amount of big data is generated from wearable devices in patients, emergency care units, etc. Structured data such as electronic patient record, health record, unstructured and semi structured data such as scanned images, clinical notes, prescriptions, signals from emergency care units, health data from social media are few examples of big data generated from health care domain. Predictive analytics on health big data helps in predicting the spread of diseases [8], [9], predicting chances of readmission in hospitals [10], predicting the diseases at an early stage, [11], [12],clinical decision support system to identify the right treatment for the affected patients, hospital management etc [13]. A detailed overview about the use of predictive analytics in health informatics is presented in [14]. The paper discusses about the applications of big data predictive analytics in health informatics, techniques for the same, opportunities and challenges.

Research gaps

Apart from storage, processing and aggregating different types of data in health care, identifying the dependencies among different data types is still an open challenge that requires optimal solution. Another challenge is with data compression methods. Though various methods are available for big data compression like FPGA, lossy image compression, they might not suit well for medical big data since medical images shouldn't lose any information [44]. An another area of improvement is in predicting the spread of diseases earlier. Though there are certain works carried out in this area, few important features were not taken into consideration for prediction. For example, though environmental attributes are used as input for predicting the spread of disease, certain inputs related to biological and sociobehavioral is excluded in the proposed approach in [8]. Proper dimensionality reduction techniques and feature selection are not considered in few works. Merely knowledge of domain experts are used for feature selection in health big data [10]. Clinical decision support systems is another challenge to work with. Though clinical decision support systems are developed, the success rate is very less. The decision support system should be developed considering both the patient's and physician's perspectives as patients' acceptance is very important for these systems. It can be of greater importance for emergency care units. Few challenges to work on include privacy issues, proper training for clinicians, quality of data etc. Such systems help in taking precautionary actions like identifying low risk patients, work on hospital readmission rates etc [60]. Radiation oncology is another area open to researchers. Building an integrative model for radiation oncology to be used as decision support system will be of great help for clinicians [10]. More concentration on genomic analysis is required since the present applications of clinical prediction uses genomic data. Research on functional path analysis of genomic data has to be concentrated upon [44]. Research works on handling noise, complexity, heterogeneity, real time, privacy in clinical big data is the need of the hour [14].

4.2 Education

Education field is another domain generating lot of big data using sensors, mobile devices for applications like learning management system, online assignments, exams, marks, attendance etc. Social media is also widely used by students and instructors as forums [15]. Predictive analytics in data generated from these devices and institutions help to predict the effectiveness of a course [15], learning method [16],[17], student's performance [18], [20], institution's performance [21] etc. Such predictions also help to personalize instructions by customizing the learning experience to each student's skills and abilities. [19] discusses about big data opportunities and challenges in education field. [17] and [21] also discusses about the scope of predictive analytics in educational big data.

Research gaps

Firstly, there are very few works carried out for predictive analytics in education field. There are very few papers in this field and most of them are review and survey papers. Hence researchers can focus on this field. The major

challenge involved in this field is social and ethical challenges. Since the student‘s and institution‘s individual data are used for prediction purposes, many students and institutions might not want their data to be exposed [68]. In the recent days, many institutions allow students to bring their own devices for language learning. Hence massive amounts of data is generated and scalability is an important area to concentrate in future [19]. Another area to work on is in integrating data from different sources. Student data are available in multiple sources like social media, schools, district offices, universities etc. Few are structured and few unstructured. A more focus on this area can help prediction.

4.3 Supply chain management

Predictive analytics helps in supply chain management. Accenture is a company that has implemented big data solutions for prediction in supply chain management [22]. Prediction in supply chain management helps to improve customer relationships by regular interactions with them thereby helping in understanding their satisfaction level, product recommendations [22], predicting supplier relationships [23], reduce the customer waiting times [24], improve the production based on demand [25], [26], manage the inventory effectively [27] and reduce risks in the process of supply chain [28].[29] discusses about the advantages of predictive big data analytics in this domain. Product development is another area where big data solutions can help the process.

Research gaps

Though there is more scope for prediction in supply chain management, very few industries have implemented it. Probably because of the hesitation to invest and the lack of skillset. Models which can reduce cost can be proposed for supply chain management processes. Hiring data scientists with domain knowledge helps the industries to move towards big data solutions for efficient supply chain management [22]. Though some of the companies use analytics in supply chain management, most of them are ad-hoc and situation specific. Predictive analysis on other areas like improvement in demand driven operations, better customer supplier relationships, optimization of inventory etc can be more concentrated upon [22], [27]. Generalized models for prediction in supply chain industry can be more focussed on. Though [23] proposed a model based on deduction graph, it is not tested on variety of product designs. Privacy of data is not considered. The approach also uses lot of mathematical techniques. Hence approaches using simple techniques can be developed. More solutions for supply chain management considering both strategy and operations has to be focussed upon [26].

4.4 Product development and marketing industry

[30] presents a white paper about the scope of predictive analytics for product development process. Marketing is a part of almost all the sectors and prediction in marketing has gained more importance because of its direct impact in business and income. Predictions using big data solutions for marketing helps to acquire customers, develop customers and retain customers. Prediction in product development and marketing industry helps to validate the design of the product, predict the demand and supply thereby increasing the sales and improving the customer experience.

Research gaps

There is no single technology available to address all the big data requirements. Big data solutions have to be integrated with other approaches and techniques to support predictive analytics. Researchers can concentrate to work on a single technology that can address all the requirements [30]. Also, different processes have to use different techniques and approaches specifically designed for them. For example,semiconductor manufacturing process should consider the spatial, temporal and hierarchical properties in manufacturing process as the existing algorithms doesn‘t suit well for them. Specific solutions can be proposed for different product development industries [55]. More work on implementing the machine learning algorithms in different areas of marketing and integrating them together can be a scope of work for researchers [45].

4.5 Transportation

‘Smart city’ is a diction that is of common talk in today‘s world. A smart city uses the information collected from sensors operating over the cities to help in the administration of the cities. Many research works are carried out in this area. Intelligent transportation systems is required for building smart city. Sensors generate lot of information that require big data solutions for processing and prediction. Predictive analytics using big data solutions for transportation has lot of applications like predicting the traffic and controlling it efficiently [31], [32], [33],predicting the travel demand and making effective use of the infrastructure thereby reducing the waiting time of the passengers [34], [35], automatic control of traffic signals [36] and predicting the transport mode of a person [37].

Research gaps

With the advent of many devices, lot of information is generated in the field of transportation from various sources. New tools to integrate data from sensors and latest devices to traditional data sources is required. Researchers can focus on developing such tools [34]. The capability of the real time traffic data collection service should be improved since video and image data are all collected [36]. Proper methods to handle correlations among data and uncertainty in data is also required in this field since the data

generated is temporal and spatial in nature for transportation. Readings from sensors are also uncertain [34]. Another area to concentrate is on using other deep learning approaches to predict traffic flow for better performance. The prediction layer used in [33] is just logistic regression. More powerful predictors can improve performance. Data fusion for social transportation data is still in a preliminary stage. GPS from taxi driver only give information about origin and destination but not on travel demand. Mobile data are used for travel demand but can't estimate travel time on roads. Hence a proper fusion approach has to be used to integrate data from several sources.Web based agent technology for transportation management and control is a research direction [35]. Software robots to monitor the state of drivers, check the condition of cars, evaluate safety environment is a research in progress [35]. There is more scope in predicting the transport mode of a person. [37] used only sensor information for classification whereas future work can concentrate on integrating information from cloud too. Advanced techniques to remove noise and outliers can be worked upon.

4.6 Other domain areas

Agriculture can be benefited using predictive analytics. Now-a-days sensors and UAVs finds its use in agriculture. Sensors are used to find the effectiveness of certain type of seed and fertilizer in different places of the farmland. Big data solutions are used to store and analyze this information to improve the operations in agriculture. Additional information like predicting the effect of certain diseases on crops helps to take precautionary measures. Farmers do predictive analytics in agricultural big data to lower costs and increase yields. The use of different fertilizers, pesticides is used to predict the environmental effects [38], [39]. Big data prediction finds its application in banking. Analysis in browsing data helps to acquire customers. Defaulters are predicted by mining Facebook data. Banks use sentiment analysis to analyze the customer needs and preferences and motivate them to buy more products thereby reducing the customer churn. Facebook interactions, tweets, customer bank visits, logs, website interactions are used as sources for sentiment analysis. A 360 degree view of customer interactions is analyzed to prevent churning of customers. Certain features like account balance, time since last transaction all helps banks to frame rules and identify the customers who are about to churn. Big data prediction helps to identify hidden customer patterns. Large sample of outliers are analyzed to predict fraud detection in banks [40]. There are also few works carried out for prediction in library big data. Libraries work with online journals and resources which are again voluminous. Predictive analytics is used in library big data for useful extractions related to learning analytics, research performance analytics etc. The user search behavior, log behavior are all analyzed to extract useful information. An other industry where predictive analytics finds its importance is telecommunication industry. Lot of applications today like Whatsapp uses different kinds of data that include structured, unstructured and semi structured. Predictive analytics in telecommunication industry focusses mainly on customer satisfaction [41]. Predictive analytics in big data helps business too. Big data analytics platforms of different providers help in personalization. The extent to which they support personalization differs in different platforms. Many big data platforms like KNIME, IBM Watson analytics are all finding its use in personalization [42]. Prediction in big data is used extensively in robotics field also. Robots communicate with many other systems. Sharing of information between robots and smart environments, comparing the information the robot has with other systems improves the robot intelligence [43]. Movie industry uses big data solutions for predicting the success using social media. Enormous data gets accumulated in social media like Wikipedia, Facebook and Twitter. Self-aware machines are finding its way in industries with the help of big data and cloud computing techniques. These machines are capable of monitoring their health condition on their own and take smart decisions on their maintenance.

5 Comprehensive challenges

Big data has its own challenges in terms of storage, processing, analytics etc. We restrict this paper to address the overall challenges involved in predictive analytics on big data and to throw light on few techniques used and state of the art work done in handling these challenges. The overall challenges are categorized under six headings shown in figure 3.

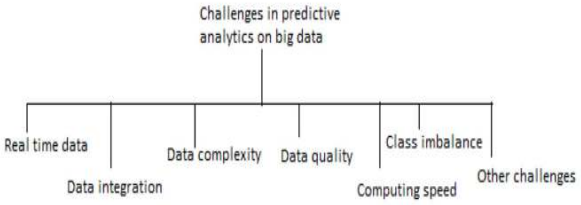

Figure 3: Comprehensive challenges of Predictive analytics on big data - taxonomy

5.1 Real-time data

Handling real time data is one among the major challenges in predictive analytics on big data. Few predictions such as predicting the early outbreak of the disease to take care of public health [9], real time recommendation system for marketing requires real time data from social sites to be collected.

Firstly, Latency is one of the main parameter to be taken care of when working with such data. Secondly, techniques

to handle interactive queries is important during predictions with these data. Thirdly, predictive algorithms should be integrated with solutions handling real time data for effective prediction.

Some big data solutions and machine learning algorithms are used in handling the above challenges with real time data. [44] states that spark and storm helps in collecting the data without latency. Hadoop platform are used to carry out Extract, transform, load(ETL) operations. Hbase, HiveQL are used to work on interactive queries. Open source software applications like Cassandra, MongoDB are used to achieve scalability and performance. Apache Mahout Machine learning library is used to run predictive algorithms on top of Hadoop [30]. Yet these techniques are generalized and their performance differs depending on the nature of the data. [72] proposes a task level adaptive mapreduce framework for real time streaming data in health care applications. The scaling capability is designed at task level that improves scalability in real time streaming data from sensors. It is proved to cope up with the velocity of the streaming data. But this approach was tested only on health care applications.

Few traditional data mining algorithms are also proposed for real time data. [44] states that algorithms such as Naive bayes can be used for sentiment analytics to extract the words from the twitter or other sites. Logistic regression, nearest neighbours are used for customer segmentation and to predict the probability that the customer will click the advertisement. [73] uses naive bayes to extract information from tweets texts. [4] proposes the use of Extreme learning machine to enhance the speed for processing real time data. [30] uses random forests and bayesian techniques to predict the crash and congestion in real time traffic monitoring. Yet, when the velocity at which the real time data enters increases, the performance of these traditional data mining algorithms deteriorate.

Mobile computing techniques and cloud infrastructure are used with big data platforms and data mining algorithms to handle real time data. [46] proposes a monitoring platform Context-Aware platform using integrated mobile services(CAPIM) to make the life of smart city easier. CAPIM collects the general traffic information, stores in the mobile device and uploads when the wi-fi is available. Drivers are provided feedback about the traffic information that helps to take decisions. More visualization output is presented to the users using google maps and the services in turn send data about his locality on his social accounts like twitter etc. [31] proposes the use of techniques like Hadoop, Hadoop Distributed File System(HDFS) and HBase to store the traffic related information. This paper proposes Real time traffic information(RTTI) system for collecting and integrating information from various sources. Massive traffic data sets are utilized transparently with the aid of cloud infrastructure. Cloud and big data is integrated in traffic flow algorithms. The usage of cloud for massive storage and the use of mapreduce techniques and Hadoop HDFS improves the performance of data mining algorithms in predictions with real time traffic flow information. Cloud services like Watson analytics is also used to analyze real time data from social media. Yet there are not many platforms that integrates cloud services, big data solutions and mobile computing. There is a need for data fusion approaches.

Few other related works for real time data include [4] that uses Representative streaming processing systems for processing real time streaming data and [46] that uses Very Fast Decision Tree (VFDT) algorithm and IBM Infosphere streams to analyze the real time streaming population data to predict the spread of cardio respiratory spells. [48] also proposes a multi-dimensional fusion technique that works on Hadoop platform with both real time and offline data. This model suits well for satellite applications where real time data is captured and forwarded to the ground station for refining and prediction. Energy efficient memrister based neural network is used for big data analytics. GPUs are used instead of CPUs to improve the speed up. Recurrent neural networks are used since they prove to be effective for non- linear sequential processing tasks like speech recognition, natural language processing, video indexing etc. [69] uses convolutional neural networks with GPU capabilities to detect real time objects. [70] combines real time measurements from real time databases with static data and uses simple extrapolation technique for prediction in substation automation. [71] proposes a compression technique for real time analog signals. It can handle the big data in real time instruments and optical communications.

5.2 Data integration

This section discusses about the challenges involved in integrating data of different types for prediction. Heterogenous data is one of the main characteristics of big data. Few predictive works require data from different sources to be integrated with the existing data. For example, predictions in educational sector collects data from multiple sources like social media, schools, district offices, universities etc. Few are structured and few are unstructured. [20] proposes a model that integrates information from social media and predicts student involvement and success. Since social media is used by students to share their ideas, feedbacks about courses, sentiment analysis can be used to solve problems by analyzing the most common feedbacks, ideas etc [16]. Healthcare sector also requires integration of information from different sources. [49] uses unstructured data from Google Flu trends to predict the spread of influenza by predicting the region that are most likely to be affected. Past data is combined with real time data for prediction. [11] proposes a predictive model to predict Systemic Lupus Erythematosus, a disease that affects multiple organs by integrating structured data like electronic health records, unstructured and semi structured data like imaging and scan tests(MRI, CT, Ultrasound scan, X-ray), complete blood count, urinalysis. [8] develops a spatial data model to predict influenza epidemic in Vellore, India. Large repos-

itories of data are collected. The developed spatial model is dependent on geographically weighted regression technique. It involves bringing together several data sources like surveillance systems, sentinel data etc to predict the spread of epidemics. Movie industry uses big data solutions for predicting the success using social media. Enormous data gets accumulated in social media like Wikipedia, Facebook and Twitter. [50] proposed a predictive model for predicting the movie box office success in Wikipedia. The major issue with heterogeneous data is that since they may not exactly be the same, there are possible chances that the machine learning results may be affected.

Big data platforms with simulation help to bring together data from different sources and predict the hidden patterns in them. [10] develops a predictive model using Mahout's machine library that works on top of Hadoop with Mapreduce technology to find out the chances of readmission after discharge of persons with heart failure. Data is collected from various sources and integrated using mahout. Mahout machine library is used for prediction that runs mapreduce jobs on Hadoop. The model uses HiveQL for distributed query. Data extraction and integration is done using Hadoop, Hive and Cassandra. Random forest and logistic regression is used in predicting the readmissions. Big data solutions gave good results in terms of time efficiency and scalability. [23] proposes a model for prediction in supply chain management process that uses deduction graph model to visually link competent sets from many data sources both structured and unstructured. Customer preferences and new product ideas are predicted through social media using recent shopping history. Customer response time is improved. Big data solutions such as Apache Mahout is used for machine learning, tableau is used for visualization, Ionosphere for data mining etc. Infrastructure proposed by combining deduction graph model with data mining, proves to provide better results in supply chain management with respect to usability, feasibility etc. [51] proposes the use Mapreduce technology in distributed data management and scheduling for heterogenous environment. The paper proposes a system for social transportation. Statistical approaches, data mining algorithms and visualization analytics are used according to the type of data. It is implemented in hadoop platform but used with other frameworks also like cascading, sailfish, Disco, Skynet, Filemap, Themis etc. High level languages like PigLatin, HiveQL are used for the technology. Hence, very few tools and techniques are available to integrate data from different sources and the advent of devices like sensors and RFIDs kindles the requirement for new tools and techniques.

Oncology can help integration easier in healthcare sector. Radiation oncology ontology is a key component used in data collecting system for better interpretability. Radiation oncology requires aggregation of input from many origins like scans, images and EHR of patients. [52] carried out a survey that explains about the state of art and future prospects of using machine learning algorithms and big data in radiation oncology. [9] states that building an integrative model for radiation oncology to be used as decision support system will be of great help for clinicians. But the research works in oncology is in its very initial stage.

Data warehousing is another concept in which data are integrated from heterogeneous sources. Few examples include [275] that proposes a dimensional warehouse for integrating data from clinical trials, [276] that proposes an architecture for a data warehouse model to integrate health data from different sources etc. The major drawback with data warehousing with big data is that technologies like hadoop, mapreduce etc. has to be integrated with the data warehouse to support the big data requirements. Moreover the ETL operations have to be designed in the architecture to suit big data requirements. There is also no standard architectural framework to design data warehouse for big data and hence the existing architectures designed to suit particular problems lacks flexibility [277]. Concepts such as data lakes are also of use as they are non relational approaches to integrate different types of data from heterogeneous resources. They postpone data mapping until query time. But, the query and reporting capabilities of data lakes are still emerging. They are not as powerful as SQL on relational databases [275]. Few research works such as [278] are proposed that manages the metadata effectively in data lakes to address the big data requirements.

Few related works on data integration for big data includes [74] for their work on a new resource 'Diseases' that aims to find the associations between genes and diseases by integrating automated text mining with existing datasets, [75] for their data integration method 'Optique' based on ontology that integrates streaming and static data as an abstraction layer, [76] for their work on API centric linked data integration that discusses about the use of API to extend the classical linked data application architecture to facilitate data integration, [78] for their work to propose a new approach for integration based on semantics. This approach converts the data sources in different formats to nested relational models initially and then imports only a subset of large datasets to build the model thereby coping up with the data size problem. A review of data integration in life sciences along with its challenges are discussed in [77]. [79] proposes the use of fused lasso approach in regression coefficients to learn heterogeneity when combining data from heterogeneous sources.

5.3 Data complexity

Large complex datasets such as genomic datasets have to be dealt in few predictive tasks. Such complex datasets make the predictive task difficult. Firstly, storage and processing of such complex data becomes a problem. Secondly, understanding such complex data to build predictive models need to be taken care of. Thirdly, either enhancements in existing data mining algorithms or new data mining algorithms have to be proposed to handle such complex datasets.

[44] states that storage techniques like HDFS, apache

Hadoop helps in storing and processing the big data effectively. Using distributed platform like mapreduce prevents the over utilization of the resources. A detailed survey on map reduce technology is discussed in [6]. Cloud platforms also help in handling complex data. Cloud platforms in health care like "PCS-on-demand"are found to be effective in storing and sharing of healthcare information with its cloud infrastructure. Mapreduce is used to parallelize the processing. Mahout library is used for machine learning algorithm to process the images, signals and genomic data. MongoDB is used for storage because of its high availability, performance and easy scalability [45]. [25] proposes the use of Microsoft Azure, a cloud platform for data storage in inventory management for supply chain process. Data sources for inventory management are internal like RFID, sensors etc. They generate lot of data and big data solutions help in processing them efficiently. NoSQL is used for data access. Batch analytics is done using Apache Hadoop. [32] proposes a model that sends driving guidance to vehicles with cloud computing technique incorporated to big data. Big data solutions [53] are used for spatial data analytics and Cloud solutions in big data platforms are used for predictive analytics in tactical big data [54].

Visualization analytics and clustering helps in understanding complex datasets. [12] develops a predictive model for diabetic data analysis in big data. Association rule mining is used to find association between the laboratory results and diabetes type of the patient. Clustering of similar patterns and classification of health risk value by patients health condition is done and predefined deductive rules are derived to predict the diabetes. The predictive model uses Hadoop/mapreduce environment for parallel processing. Visualization also helps in predicting hidden insights from the data. [13] proposes a model for hospital management. The temporal information helps to understand the clinical activities. Proper visualization and clustering of this temporal data helps to understand the abnormalities. [55] proposes an optimization framework for wafer quality prediction in semiconductor manufacturing process that uses clustering to identify similar behavior pattern over time for chambers. [56] presents a survey on clustering time series data. Abnormality can also be discovered thereby helping quality control and fault diagnostics. But since time series clustering is mostly for unstructured data, a co-clustering pattern is formulated for this problem with constraints to match the tools and the chambers. Visualizing the data effectively helps in prediction. Aggregation and multi-dimensional analysis is also used in big data to extract knowledge from them. AsterixDB, DGFIndex are used that helps in aggregation and multi-dimensional analysis for big data [57]. Yet, new frameworks for visualization techniques and multi-dimensional analysis need to be developed exclusively for big data.

Few data mining algorithms are used in complex datasets such as genomic data and signals. [47] uses Nearest Centroid Based Classifier (NCBC) to predict clinical outcome related to cancer using gene expression data.[58] uses Multipartiate graph for prediction in genomics big data. Deep learning also helps in handling complex data. Lot of research works are carried out in clinical image processing with deep learning. [33] uses deep learning for predicting traffic flow. A stacked autoencoder model trained greedily learns traffic flow features. It uses spatial and temporal correlations.The medical data including cardiac MRI involves signal processing. There are many statistical learning tools for signal processing. [59] uses signal processing techniques such as kernel based interpolators and timely matrix decompositions for big data. Yet, more research works are required for signal processing with big data and on genomic data analysis.

Few related works to handle complex data include [80] for their work on anytime density based clustering for complex data to improve scalability, [81] for their work on interactive data visualization to understand complex data, [83] for their work to propose a Pairwise weighted ensemble clustering algorithm to cluster complex data for better understanding, [82] for their work to address scalability problem in complex data by proposing two suboptimal algorithms to address casting complex problem of L1 Norm principal component analysis, [84] for their work to develop complexheatmap package that helps in visualizing and revealing the patterns in high dimensional genomic data.

5.4 Data quality

Low quality data is another challenge in predictive analytics. The source data in some applications like from emergency care, sensors may be of poor quality. Predictive analytics techniques on such low quality data need some sophisticated techniques to be applied on the existing algorithms. Low quality data may also be due to the fact that some predictive works fail to consider important attributes required for prediction. For example, [8] predicts the spread of the disease using environmental attributes but fail to consider the biological and socio-behavioral attributes. Data may also be incomplete and inconsistent in certain cases.

Generally techniques like Mathematical or logical regression might work for low quality data [60]. [4] states that advanced deep learning methods, statistical learning theory of sparse matrix are used to overcome the challenge of incomplete and inconsistent data. Techniques like Watson analytics are used to overcome the challenge of low value density data. More works on identifying proper correlations among inconsistent data is required.

Proper dimensionality reduction techniques and feature selection also help to improve the data quality. [9] states that merely knowledge of domain experts are used for feature selection in most of the predictive works.

The nature of data differs depending on the application. For example,semiconductor manufacturing process should consider the spatial, temporal and hierarchical properties in manufacturing process as the existing algorithms doesn't

suit well for them. Specific solutions can be proposed for different domains depending on the nature of the data [54].

Few related works on predictive analytics with low quality data include [85] that proposes an extension to likelihood method to handle low quality data, [86] to propose a method that uses data mining tasks such as clustering to extract patterns from noisy data in market segmentation, [87] to propose a new algorithm based on C4.5 decision tree that uses imprecise probabilities in classifying noisy data, [88] that proposes a new algorithm for extreme machine learning to work efficiently in the presence of outliers. [89] proposes a hybrid feature selection scheme to reduce the performance deterioration caused by outliers in predictive analytics.

5.5 Computing speed

Computing speed is one of the important challenges to be handled during predictions on big data. Most of the wearable devices consume more power and the algorithms used on them are computationally intensive. Computing speed of predictive algorithms on big data also increases due to its volume.

Parallel computing techniques help to overcome the challenge with respect to volume and computing speed. [4] proposes the use of alternating direction method of multipliers to overcome the challenge with respect to volume since it acts as a platform for distributed frameworks with parallel computing. Mapreduce is used to work parallel on the chunks of the big data. [63] proposes a data mining algorithm K Nearest neighbor based on Spark(KNN-IS) for classification in big data. The algorithm uses Mapreduce technology for parallel processing of the training data set. Though Hadoop works well with mapreduce, it has its own limitations like latency which is overcome by spark's in-memory computations. Map reduce is used on spark for KNN to yield better results in terms of time and accuracy. Resilient distributed databases are used on spark platform. Sometimes medium quality predictions with low latency perform better than high quality predictions with more latency. [90] proposes parallel random forest algorithm in spark cloud computing platform to improve the computational efficiency of big data analytics. A parallel version of deep neural network training is proposed in [91].

Feature selection techniques like Representation learning, deep learning and dimensionality reduction are also used to reduce the computing speed since the unnecessary features are eradicated. Yet, the computational complexity of certain feature selection techniques like wrapper approaches is high and researchers are working on it. [92] proposes a hierarchical attribute reduction algorithm using mapreduce in which attribute reduction process is executed in parallel. [93] proposes fast minimum redundancy maximum relevance algorithm for feature selection in high dimensional data.

Fuzzy techniques aid in reducing the processing time. Researchers worked towards reducing the time for processing using fuzzy rules on data with lot of input features. An algorithm named Chi-Fuzzy rule based classification systems(Chi-FRBCS) is proposed. It works on Mapreduce framework and uses linguistic fuzzy rules. Two versions of Chi-FRBCA algorithms are proposed - Chi-FRBCS BigData-Max and Chi-FRBCS BigData-Ave. Experiments are conducted on six different big data problems set and Chi-FRBCS is proved to be effective in terms of processing time and accuracy [61]. [62] proposed a big data algorithm called FMM based on fuzzy rules for sentiment analysis in social media. A parallelized algorithm FMM with mapreduce is also used and that proves to be effective in terms of accuracy compared to the other techniques. The algorithm is made to work on twitter data and is observed that the execution time is much lesser for big data.

[64] states that the hardware solution is effective in terms of energy savings, power efficiency. Scientific applications use multi-dimensional data sets. Processing has to be faster compared with other applications because of the velocity at which it arrives. For example, predicting climate change requires fast processing. [65] proposes the use of an I/O in-memory server for scientific big data analytics applications. [37] proposes SVM polynomial degree 3 kernel to reduce the computational complexity and memory requirement during classification. This model detects the transport mode of a person whether he or she is walking, jogging or going in bike. Hardware changes are also done by using virtual gyroscope, accelerometer and few other hardware devices to ensure that low power consuming devices are used. The memory consumed by the algorithm is less.

5.6 Class imbalance

Class imbalance is another problem in certain predictive works. Techniques like oversampling and undersampling are used in class imbalance problems.

Some machine learning algorithms are effective in solving class imbalance problems. [66] proposes an algorithm ROSEFW-RF for contact map prediction. It is a classification task related to protein structure where there are very few positive samples available. This algorithm is based on key-value pairs and uses mapreduce approach for distributed processing. Predictive model is constructed using random forest. The class distribution is balanced through random oversampling. Irrelevant features are removed through feature weighing. Oversampling is found to be more robust than undersampling and cost sensitive learning when number of maps is increased in mapreduce technique. The test data is classified. Experiments are conducted in bioinformatics data and ROSEFW-RF algorithm is the winner algorithm for imbalance big data classification problem. [67] proposes Random forest with Mapreduce for prediction on imbalanced big data. Five different versions of random forest algorithms are used in imbalanced big data classification with Mapreduce approach. - RF - BigData, RF-BigDataCS, ROS+RF-BigData, RUS+RF-

BigData, SMOTE+RF-BigData. Random forest techniques is found to work well for imbalanced big data classification.

Few other related works on class imbalance problem include [94] that proposes an ensemble method to handle both online learning and imbalance learning using oversampling and undersampling bagging techniques, [95] that proposes a diversity based oversampling approach to create new instances for minority class, [96] that proposes a new ensemble method 'hardensemble' to handle class imbalance, [97] that uses extreme learning machine to handle class imbalance problem both at feature and algorithmic levels.

Class imbalance problem is another area that require researcher's attention. A mixed strategy of oversampling and undersampling can be tried to boost performance [66].

5.7 Other challenges

Few other challenges include privacy issues, lack of proper training to the domain experts, social and ethical challenges etc. For example, the decision support system in healthcare should be developed considering both the patient's and physician's perspectives as patients' acceptance is very important for these systems. It can be of greater importance for emergency care units. Privacy issues, proper training for clinicians, quality of data are all few issues to be considered in this case [60]. In education sector, since the student‘s and institution‘s individual data are used for prediction purposes, many students and institutions might not want their data to be exposed [68]. Hiring data scientists with domain knowledge helps the industries to move towards big data solutions for efficient supply chain management [22].

6 Potential research directions

From the above discussions on predictive analytics with big data, it has been observed that this field is readily open for researchers. The potential research directions are summarized.

6.1 Data management

- Since real time data includes the collection of lot of video and image data in the present scenario, there is a need to improve the real time data collection service. Researchers can work on the same.
- Developing framework fusion approaches to integrate big data solutions, cloud computing techniques and mobile computing techniques is another area for research as there is no single technology available to address all the big data requirements. Researchers can concentrate to work on a single technology that can address all the requirements to support predictive analytics.
- With the advent of many devices, lot of information is generated in different domains from various sources. New tools to integrate data from latest devices to existing data sources is required. Researchers can focus on developing such tools.
- Data partitioning methods, indexing and multidimensional analysis on big data are few other topics for researchers.
- Scalability is an important area to concentrate in future as massive amounts of data is generated.

6.2 Algorithms and solutions

- Enhancements in traditional data mining algorithms to handle and analyze real time data or developing new algorithms for the same is another promising research area available for researchers.
- Identifying the dependencies and semantic features among heterogeneous data types is still an open challenge requiring favorable solutions due to the biased view of data distribution.
- New visualization techniques and frameworks can be developed for effective interpretation of complex data.
- Genomic data analytics is in its initial stage. Works such as functional path analysis on genomic data is an area open for researchers.
- Data compression methods is also a challenge. Big data compression methods like FPGA, lossy image compression, might not suit well for certain types of big data like medical images as they shouldn‘t lose any information. Hence there is a need for new data compression methods exclusively for specific big data types.
- Oncology, semantic web are all areas to be concentrated in machine learning for big data.
- Establishing correlations among uncertain data, temporal and spatial data is another area to work on for researchers.
- Overfitting still remains as an open issue and researchers can focus on developing better solutions without much compromise on accuracy and cost.
- Using ensemble techniques in mapreduce platform is another area that can be concentrated on to improve accuracy
- Specific solutions can be proposed for different domains depending on the nature of the data.
- Class imbalance problem is another area that require researcher's attention. A mixed strategy of oversampling and undersampling can be tried to boost performance

7 Conclusion

The paper provides an overview of core predictive models and their challenges on big data. We discussed the scope of predictive analytics on big data generated across different domains and few research gaps are identified. Though the mathematical approaches may not suit well for big data, we found that the data mining approaches and machine learning techniques used for prediction have their base from the mathematical approaches. The choice of predictive model depends on the nature of application and data in hand. Finally we presented comprehensive challenges of predictive analytics on big data and state of the art techniques used to address the challenges. Based on our discussion, we also presented a separate section on future directions for research.

References

[1] Chiang, Roger H.L and Goes, Paulo and Stohr, Edward A (2012) Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline, *ACM Transactions on Management Information Systems (TMIS)*, https://doi.org/10.1145/2361256.2361257.

[2] Jain, A. K. and Murty, M. N. and Flynn, P. J (1999) Data Clustering: A Review, *ACM Comput. Surv.*, 264–323, https://doi.org/10.1145/331499.331504.

[3] Benjamins, Richard.V (2014) Big Data: from Hype to Reality?, *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), ACM*, https://doi.org/10.1145/2611040.2611042.

[4] Qiu, Junfei and Wu, Qihui and Xu, Yuhua and Feng, Shuo (2016) A survey of machine learning for big data processing, *EURASIP Journal on Advances in Signal Processing*, https://doi.org/10.1186/s13634-016-0382-7.

[5] Chen, Ju-Chin and Liu, Chao-Feng (2015) Visual-based Deep Learning for Clothing from Large Database, *Proceedings of the ASE BigData and SocialInformatics, ACM*.

[6] Sakr, Sherif and Liu, Anna and Fayoumi, Ayman G (2013) The Family of Mapreduce and Large-scale Data Processing Systems, *ACM Comput. Surv., ACM*, https://doi.org/10.1201/b17112-3.

[7] Hung, San-Chuan and Kuo, Tsung-Ting and Lin, Shou-De (2015) Novel Topic Diffusion Prediction Using Latent Semantic and User Behavior,*Proceedings of the ASE BigData and SocialInformatics, ACM*.

[8] D. Lopez and M. Gunasekaran and B. S. Murugan and H. Kaur and K. M. Abbas (2014) Spatial big data analytics of influenza epidemic in Vellore, India,*IEEE International Conference on Big Data*, https://doi.org/10.1109/BigData.2014.7004422.

[9] Curran, Martina and Howley, Enda and Duggan, Jim (2016) An Analytics Framework to Support Surge Capacity Planning for Emerging Epidemics,*Proceedings of the 6th International Conference on Digital Health Conference, ACM*, https://doi.org/10.1145/2896338.2896354.

[10] Zolfaghar, Kiyana and Meadem, Naren and Teredesai, Ankur and Roy, Senjuti Basu and Chin, Si-Chi and Muckian, Brian (2013) Big data solutions for predicting risk-of-readmission for congestive heart failure patients,*IEEE International Conference on Big Data*, https://doi.org/10.1109/BigData.2013.6691760.

[11] S. Gomathi and V. Narayani (2015) Implementing Big Data analytics to predict Systemic Lupus Erythematosus,*International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, https://doi.org/10.1109/ICIIECS.2015.7192893.

[12] Eswari, T and Sampath, P and Lavanya, S and others (2015) Predictive Methodology for Diabetic Data Analysis in Big Data,*Procedia Computer Science*, pp:203–208, https://doi.org/10.1016/j.procs.2015.04.069

[13] Tsumoto, Shusaku and Hirano, Shoji (2015) Analytics for Hospital Management,*Proceedings of the ASE BigData and SocialInformatics, ACM*.

[14] Fang, Ruogu and Pouyanfar, Samira and Yang, Yimin and Chen, Shu-Ching and Iyengar, S.S (2016) Computational health informatics in the big data age: a survey,*ACM Computing Surveys (CSUR)*, pp.12, https://doi.org/10.1145/2932707.

[15] The center for Digital education (2015) : Big data in education report. Technical Report, https://blog.stcloudstate.edu.

[16] Sin, Katrina and Muthu, Loganathan (2015) Application of big data in education data mining and learning analytics?A literature review,*ICTACT Journal on Soft Computing*, pp.1–035, https://doi.org/10.21917/ijsc.2015.0145.

[17] Oracle (2015) Big data education. Technical Report, www.oracle.com.

[18] Jo, Il-Hyun and Kim, Dongho and Yoon, Meehyun (2014) Analyzing the Log Patterns of Adult Learners in LMS Using Learning Analytics,*Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, ACM*, pp.183–187, https://doi.org/10.1145/2567574.2567616.

[19] Wang, Yinying (2016) Big opportunities and big concerns of big data in education,*TechTrends, Springer*, pp.1–4 , https://doi.org/10.1007/s11528-016-0072-1.

[20] Niemi, Gitin and David, Elena (2012) Using Big Data to Predict Student Dropouts: Technology Affordances for Research,*International Association for Development of the Information Society*.

[21] Kellen, V and Consortium, Cutter and Recktenwald, A and Burr, S (2013) Applying big data in higher education: A case study,*Data Insight and Social BI*.

[22] Accenture (2014) Accenture-Global-Operations-Megatrends-Study-Big-Data-Analytics. Technical Report, https://acnprod.accenture.com.

[23] Tan, Kim Hua and Zhan, YuanZhu and Ji, Guojun and Ye, Fei and Chang, Chingter (2015) Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph,*International Journal of Production Economics*, pp:223–233, https://doi.org/10.1016/j.ijpe.2014.12.034.

[24] Rozados,Ivan Varela and Tjahono, Benny (2014) Big Data Analytics in Supply Chain Management: Trends and Related Research,*6th International Conference on Operations and Supply Chain Management*.

[25] J.Leveling and M. Edelbrock and B. Otto (2014) Big data analytics for supply chain management,*IEEE International Conference on Industrial Engineering and Engineering Management*, pp:918-922, https://doi.org/10.1109/IEEM.2014.7058772.

[26] Wang, Gang and Gunasekaran, Angappa and Ngai, Eric WT and Papadopoulos, Thanos (2016) Big data analytics in logistics and supply chain management: Certain investigations for research and applications,*International Journal of Production Economics, Elsevier*, pp:98–110, https://doi.org/10.1016/j.ijpe.2016.03.014.

[27] Waller, Matthew A and Fawcett, Stanley E (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management,*Journal of Business Logistics*, pp:77–84, https://doi.org/10.1111/jbl.12010.

[28] Wilschut, Tim and Adan, Ivo JBF and Stokkermans, Joep (2014) Big data in daily manufacturing operations,*Proceedings of the Winter Simulation Conference*, pp:2364–2375, https://doi.org/10.1109/WSC.2014.7020080.

[29] He, Miao and Ji, Hao and Wang, Qinhua and Ren, Changrui and Lougee, Robin (2014) Big data fueled process management of supply risks: sensing, prediction, evaluation and mitigation,*Proceedings of the Winter Simulation Conference*, pp:1005–1013, https://doi.org/10.1109/WSC.2014.7019960.

[30] Intel (2013) Predictive analytics and interactive queries on big data.Technical Report, https://software.intel.com.

[31] Shi, Qi and Abdel-Aty, Mohamed (2015) Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways,*Transportation Research Part C: Emerging Technologies*, pp:380–394, https://doi.org/10.1016/j.trc.2015.02.022.

[32] Yu, Jianjun and Jiang, Fuchun and Zhu, Tongyu (2013) RTIC-C: a big data system for massive traffic information mining,*International Conference on Cloud Computing and Big Data (CloudCom-Asia), IEEE*, pp:395–402.

[33] Y. Lv and Y. Duan and W. Kang and Z. Li and F. Y. Wang (2015) Traffic Flow Prediction With Big Data: A Deep Learning Approach,*IEEE Transactions on Intelligent Transportation Systems*, pp:865-873.

[34] Toole, Jameson L and Colak, Serdar and Sturt, Bradley and Alexander, Lauren P and Evsukoff, Alexandre and Gonzalez, Marta C (2015) The path most traveled: Travel demand estimation using big data resources,*Transportation Research Part C: Emerging Technologies*, pp:162–177, https://doi.org/10.1016/j.trc.2015.04.022.

[35] Yuan, Nicholas Jing and Zheng, Yu and Zhang, Liuhang and Xie, Xing (2013) T-finder: A recommender system for finding passengers and vacant taxis,*IEEE Transactions on Knowledge and Data Engineering*, pp:2390–2403, https://doi.org/10.1109/TKDE.2012.153.

[36] Wang, Chao and Li, Xi and Zhou, Xuehai and Wang, Aili and Nedjah, Nadia (2016) Soft computing in big data intelligent transportation systems,*Applied Soft Computing, Elsevier*, pp:1099–1108, https://doi.org/10.1016/j.asoc.2015.06.006.

[37] Yu, Meng-Chieh and Yu, Tong and Wang, Shao-Chen and Lin, Chih-Jen and Chang, Edward Y (2014) Big data small footprint: the design of a low-power classifier for detecting transportation modes,*Proceedings of the VLDB Endowment, VLDB Endowment*, pp:1429–1440, https://doi.org/10.14778/2733004.2733015.

[38] Stubbs, Megan (2016) Big Data in U.S Agriculture.Technical Report.

[39] Sangwani, G (2016) The Big Future of Big Data,*Business Insider, India*.Technical Report.

[40] Martens (2016) Financial forum.Technical Report, https://www.financialforum.be.

[41] Malaka, Iman and Brown, Irwin (2015) Challenges to the Organisational Adoption of Big Data Analytics: A Case Study in the South African Telecom-

munications Industry,*Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists, ACM*, pp:27:1–27:9, https://doi.org/10.1145/2815782.2815793.

[42] Lopes, claudio and Cabral, Bruno and Bernardino, Jorge (2016) Personalization Using Big Data Analytics Platforms,*Proceedings of the Ninth International Conference on Computer Science & Software Engineering, ACM*, pp:131–132.

[43] Felzmann, Heike and Beyan, Timur and Ryan, Mark and Beyan, Oya (2016) Implementing an Ethical Approach to Big Data Analytics in Assistive Robotics for Elderly with Dementia,*SIGCAS Comput. Soc., ACM*, pp:280–286, https://doi.org/10.1145/2874239.2874279.

[44] Belle, Ashwin and Thiagarajan, Raghuram and Soroushmehr, S.M.Reza and Navidi, Fatemeh and A.Beard, Daniel and Najarian, Kayvan (2015) Big data analytics in healthcare,*BioMed research international*, https://doi.org/10.1155/2015/370194.

[45] EVRY (2014) Big data - white paper.Technical Report, www.evry.com.

[46] Dobre, C and Xhafa, F (2014) Intelligent services for big data science,*Future Generation Computer Systems*, pp:267–281, https://doi.org/10.1016/j.future.2013.07.014.

[47] Herland, Matthew and Khoshgoftaar, Taghi M. and Wald, Randal (2014) A review of data mining using big data in health informatics,*Journal Of Big Data*, pp:1–35, https://doi.org/10.1186/2196-1115-1-2.

[48] Ahmad, Aswais and Paul, Anand and Rathore, Mazhar and Chang, Hangbae (2016) An efficient multidimensional big data fusion approach in machine-to-machine communication,*ACM Transactions on Embedded Computing Systems (TECS)*.

[49] Davidson, Michael W and Haim, Dotan A. and Radin, Jennifer M (2015) Using networks to combine big data and traditional surveillance to improve influenza predictions,*Scientific reports*, https://doi.org/10.1038/srep08154.

[50] Mestyan, Marton and Yasseri, Taha and Kertesz, Janos (2013) Early prediction of movie box office success based on Wikipedia activity big data,*PloS one*, https://doi.org/10.1371/journal.pone.0071226.

[51] Zheng, Xinhu and Chen, Wei and Wang, Pu and Shen, Dayong and Chen, Songhang and Wang, Xiao and Zhang, Qingpeng and Yang, Liuqing (2016) Big data for social transportation, *IEEE Transactions on Intelligent Transportation Systems*, pp:620–630, https://doi.org/10.1109/TITS.2015.2480157.

[52] Bibault,Jean Emmanuel and Giraud, Philippe and Burgun, Anita (2016) Big Data and machine learning in radiation oncology: State of the art and future prospects,*Cancer letters*, https://doi.org/10.1016/j.canlet.2016.05.033.

[53] Chen, Xin and Vo, Haong and Aji, Ablimit and Wang, Fusheng (2014) High performance integrated spatial big data analytics,*Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, ACM*, pp. 11–14.

[54] Savas, Onur and Sagduyu, Yalin and Deng, Julia and Li, Jason (2014) Tactical Big Data Analytics: Challenges, Use Cases, and Solutions,*SIGMETRICS Perform. Eval. Rev., ACM*, pp.86–89, https://doi.org/10.1145/2627534.2627561.

[55] Zhu, Yada and Xiong, Jinjun (2015) Modern Big Data Analytics for Old-fashioned Semiconductor Industry Applications,*Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp.776–780.

[56] Liao, T Warren (2005) Clustering of time series data survey,*Pattern recognition*, pp.1857–1874, https://doi.org/10.1016/j.patcog.2005.01.025.

[57] Cuzzocrea, Alfredo (2015) Aggregation and multidimensional analysis of big data for large-scale scientific applications: models, issues, analytics, and beyond,*Proceedings of the 27th International Conference on Scientific and Statistical Database Management, ACM*, pp. 23, https://doi.org/10.1145/2791347.2791377.

[58] Phillips, Charles A. and Wang, Kai and Bubier, Jason and Baker, Erich J. and Chesler, Elissa J. and Langston, Michael A. (2015) Scalable Multipartite Subgraph Enumeration for Integrative Analysis of Heterogeneous Experimental Functional Genomics Data,*Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, ACM*, pp.626–633, https://doi.org/10.1145/2808719.2812595.

[59] Slavakis, Konstantinos and Giannakis, Georgios B and Mateos, Gonzalo (2014) Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge,*IEEE Signal Processing Magazine*, pp.18–31, https://doi.org/10.1109/MSP.2014.2327238.

[60] Alexander T. Janke and Daniel L. Overbeek and Keith E. Kocher and Phillip D. Levy (2016) Exploring the Potential of Predictive Analytics and Big Data in Emergency Care,*Annals of Emergency Medicine*, pp.227 - 236, https://doi.org/10.1016/j.annemergmed.2015.06.024.

[61] Victoria Lopez and Sara del Rio and Jose Manuel Benitez and Francisco Herrera (2015) Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data,*Fuzzy Sets and Systems*, pp.5 - 38, https://doi.org/10.1016/j.fss.2014.01.015.

[62] Bing, Li and Chan, Keith C.C (2014) A fuzzy logic approach for opinion mining on large scale twitter data,*Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing* , pp. 652–657.

[63] Jesus Maillo and Sergio Ramirez and Isaac Triguero and Francisco Herrera (2016) kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data,*Knowledge-Based Systems*.

[64] Wang, Yu and Li, Boxun and Luo, Rong and Chen, Yiran and Xu, Ningyi and Yang, Huazhong (2014) Energy efficient neural networks for big data analytics,*Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp:1–2.

[65] Elia, Donatello and Fiore, Sandro and D'Anca, Alessandro and Palazzo, Cosimo and Foster, Ian and Williams, Dean N (2016) An in-memory based framework for scientific data analytics,*Proceedings of the ACM International Conference on Computing Frontiers, ACM*, pp. 424–429.

[66] Triguero, Isaac and del Rio, Sara and Lopez, Victoria and Bacardit, Jaume and Benitez, Jose M and Herrera, Francisco (2015) ROSEFW-RF: the winner algorithm for the ECBDL14 big data competition: an extremely imbalanced big data bioinformatics problem,*Knowledge-Based Systems*, pp.69–79, https://doi.org/10.1016/j.knosys.2015.05.027.

[67] Sara del Rio and Victoria Lopez and Jose Manuel Benitez and Francisco Herrera (2014) On the use of MapReduce for imbalanced big data using Random Forest,*Information Sciences*, pp.112 - 137, https://doi.org/10.1016/j.ins.2014.03.043.

[68] Johnson, Jeffrey Alan (2014) The ethics of big data in higher education,*International Review of Information Ethics*, pp.4–9.

[69] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in Neural Information Processing Systems 28*, pp.91–99.

[70] S. S. Biswas, A. K. Srivastava and D. Whitehead (2015) A Real-Time Data-Driven Algorithm for Health Diagnosis and Prognosis of a Circuit Breaker Trip Assembly, *IEEE Transactions on Industrial Electronics, vol. 62, no. 6*, pp. 3822-3831, https://doi.org/10.1109/TIE.2014.2362498.

[71] Mohammad H. Asghari and Bahram Jalali (2014) Experimental demonstration of optical real-time data compression, *Applied Physics Letters*.

[72] Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li, Kai Hwang (2015) A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications, *Future Generation Computer Systems, Volumes 43-44*, pp. 149-160, https://doi.org/10.1016/j.future.2014.06.009.

[73] Yiming Gu, Zhen (Sean) Qian, Feng Chen (2016) From Twitter to detector: Real-time traffic incident detection using social media data,*Transportation Research Part C: Emerging Technologies, Volume 67*, pp. 321-342, https://doi.org/10.1016/j.trc.2016.02.011.

[74] Sune Pletscher-Frankild, Albert Palleja, Kalliopi Tsafou, Janos X. Binder, Lars Juhl Jensen (2015) DISEASES: Text mining and data integration of disease?gene associations,*Methods, Volume 74*, pp. 83-89, https://doi.org/10.1016/j.ymeth.2014.11.020.

[75] Evgeny Kharlamov, Sebastian Brandt, Ernesto Jimenez, Ruiz, Yannis Kotidis, Steffen Lamparter, Theofilos Mailis, Christian Neuenstadt, Ozgur L. Ozcep, Christoph Pinkel, Christoforos Svingos, Dmitriy Zheleznyakov, Ian Horrocks, Yannis E. Ioannidis and Ralf Moller (2016) Ontology-Based Integration of Streaming and Static Relational Data with Optique,*Proc. of International Conference on Management Data (SIGMOD)*, pp.2109–2112, https://doi.org/10.1145/2882903.2899385.

[76] Paul Groth, Antonis Loizou, Alasdair J.G. Gray, Carole Goble, Lee Harland, Steve Pettifer (2014) API-centric Linked Data integration: The Open PHACTS Discovery Platform case study,*Web Semantics: Science, Services and Agents on the World Wide Web, Volume 29*, pp. 12-18.

[77] Gomez-Cabrero, David and Abugessaisa, Imad and Maier, Dieter and Teschendorff, Andrew and Merkenschlager, Matthias and Gisel, Andreas and Ballestar, Esteban and Bongcam-Rudloff, Erik and Conesa, Ana and Tegner, Jesper (2014) Data integration in the era of omics: current and future challenges, *BMC Systems Biology*, https://doi.org/10.1186/1752-0509-8-S2-I1.

[78] Craig A. Knoblock, Pedro Szekely (2015) Exploiting Semantics for Big Data Integration, *Association for the Advancement of Artificial Intelligence*.

[79] Lu Tang, Peter X.K. Song (2016) Fused Lasso Approach in Regression Coefficients Clustering Learning Parameter Heterogeneity in Data Integration, *Journal of Machine Learning Research*.

[80] Son T. Mai, Xiao He, Jing Feng, Claudia Plant, Christian Bohm (2016) Anytime density-based clustering of complex data, *Knowledge and Information Systems*, pp 319-355.

[81] Diane J. Janvrin, Robyn L. Raschke, William N. Dilla (2014) Making sense of complex data using interactive data visualization,*Journal of Accounting Education, Volume 32, Issue 4*, pp 31-48, https://doi.org/10.1016/j.jaccedu.2014.09.003.

[82] N. Tsagkarakis, P. P. Markopoulos, G. Sklivanitis and D. A. Pados (2018) L1-Norm Principal-Component Analysis of Complex Data, *IEEE Transactions on Signal Processing, vol. 66, no. 12, 15 June15*, pp. 3256-3267, https://doi.org/10.1109/TSP.2018.2821641.

[83] Vladimir Berikov (2014) Weighted ensemble of algorithms for complex data clustering,*Pattern Recognition Letters, Volume 38*, pp 99-106, https://doi.org/10.1016/j.patrec.2013.11.012.

[84] Zuguang Gu Roland Eils Matthias Schlesner (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data,*Bioinformatics, Volume 32, Issue 18*, pp 2847-2849, https://doi.org/10.1093/bioinformatics/btw313.

[85] Thierry Denoeux (2014) Likelihood-based belief function: justification and some extensions to low-quality data ,*International Journal of Approximate Reasoning*, pp.1535-1547, https://doi.org/10.1016/j.ijar.2013.06.007.

[86] Paul W. Murray, Bruno Agard, Marco A. Barajas (2017) Market segmentation through data mining: A method to extract behaviors from a noisy data set,*Computers and Industrial Engineering, Volume 109*, pp 233-252, https://doi.org/10.1016/j.cie.2017.04.017.

[87] Carlos J. Mantas, Joaquin Abellan : Credal-C4.5 (2014) Decision tree based on imprecise probabilities to classify noisy data,*Expert Systems with Applications, Volume 41, Issue 10*, pp.4625-4637, https://doi.org/10.1016/j.eswa.2014.01.017.

[88] B. Frenay and M. Verleysen (2016) Reinforced Extreme Learning Machines for Fast Robust Regression in the Presence of Outliers, *IEEE Transactions on Cybernetics, vol. 46, no. 12*, pp. 3351-3363, https://doi.org/10.1109/TCYB.2015.2504404.

[89] M. Kang, M. R. Islam, J. Kim, J. Kim and M. Pecht (2016) A Hybrid Feature Selection Scheme for Reducing Diagnostic Performance Deterioration Caused by Outliers in Data-Driven Diagnostics,*IEEE Transactions on Industrial Electronics, vol. 63, no. 5*, pp.3299-3310, https://doi.org/10.1109/TIE.2016.2527623.

[90] J. Chen et al (2017) A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment,*IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4*, pp. 919-933, https://doi.org/10.1109/TPDS.2016.2603511.

[91] I. Chung et al (2017) Parallel Deep Neural Network Training for Big Data on Blue Gene/Q, *IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 6*, pp. 1703-1714, https://doi.org/10.1109/TPDS.2016.2626289.

[92] Jin Qian, Ping Lv, Xiaodong Yue, Caihui Liu, Zhengjun Jing (2015) Hierarchical attribute reduction algorithms for big data using MapReduce,*Knowledge-Based Systems, Volume 73*, pp. 18-31, https://doi.org/10.1016/j.knosys.2014.09.001.

[93] Sergio Ramirez,Gallego,Iago Lastra, David Martinez,Rego, Veronica Bolon,Canedo, Jose Manuel Benitez,Francisco Herrera, Amparo Alonso,Betanzos (2016) Fast,mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High Dimensional Big Data,*International Journal of Intelligent Systems*, https://doi.org/10.1002/int.21833.

[94] S. Wang, L. L. Minku and X. Yao (2015) Resampling-Based Ensemble Methods for Online Class Imbalance Learning, *IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5*, pp. 1356-1368, https://doi.org/10.1109/TKDE.2014.2345380.

[95] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden and S. Mensah (2018) MAHAKIL- Diversity Based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction, *IEEE Transactions on Software Engineering, vol. 44, no. 6*, pp. 534-550, https://doi.org/10.1109/TSE.2017.2731766.

[96] Loris Nanni, Carlo Fantozzi, Nicola Lazzarini (2015) Coupling different methods for overcoming the class imbalance problem,*Neurocomputing, Volume 158*, pp. 48-61, https://doi.org/10.1016/j.neucom.2015.01.068.

[97] Zhen Liu, Deyu Tang, Yongming Cai, Ruoyu Wang, Fuhua Chen (2017) A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data, *Neurocomputing, Volume 266*, pp. 641-650, https://doi.org/10.1016/j.neucom.2017.05.066.

[98] Tavish Srivastava (2015) Perfect way to build a Predictive Model, Analytics Vidhya, https://www.analyticsvidhya.com.

[99] Jeremy Howick, Paul Glasziou, Jeffrey K. Aronson (2013) Problems with using mechanisms to solve the problem of extrapolation, *Theoretical Medicine and Bioethics, Volume 34, Issue 4*, pp 275–291, https://doi.org/10.1007/s11017-013-9266-0.

[100] Manuel Martin-Flores, Monique D. Parr, Luis Campoy, Robin D. Gleed (2012) The sensitivity of sheep to vecuronium: an example of the limitations of extrapolation, *Canadian Journal of Anesthesia, Volume 59, Issue 7*, pp 722–723, https://doi.org/10.1007/s12630-012-9707-7.

[101] James R. Miller, Monica G. Turner, Erica A. H. Smithwick, C. Lisa Dent, Emily H. Stanley (2004) Spatial Extrapolation: The Science of Predicting Ecological Patterns and Processes, *BioScience, Volume 54, Issue 4*,Pages 310–320, https://doi.org/10.1641/0006-3568(2004)054[0310:SETSOP]2.0.CO;2.

[102] Forbes, V. E., Calow, P. and Sibly, R. M. (2009) The extrapolation problem and how population modeling can help, *Environmental Toxicology and Chemistry*, pp: 1987-1994, https://doi.org/10.1897/08-029.1.

[103] Stack Exchange (2016) What is wrong with Extrapolation, *StackExchange*, https://stats.stackexchange.com/questions.

[104] Peter Flom (2018) The disadvantages of linear regression, *Sciencing*, https://sciencing.com/disadvantages-linear-regression-8562780.html.

[105] Benjamin A. Goldstein, Ann Marie Navar, Rickey E. Carter (2017) Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges, *European Heart Journal, Volume 38, Issue 23*, Pages 1805–1814.

[106] Harvey J Motulsky and Ronald E Brown (2006) Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate, *BMC Bioinformatics*.

[107] Johansen, S., and Nielsen, B (2016) Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models. *Scand J Statist*, 43: 321– 348, https://doi.org/10.1111/sjos.12174.

[108] Minitab Blog Editor (2013) Enough Is Enough! Handling Multicollinearity in Regression Analysis, *The Minitab Blog*, https://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in\-regression-analysis.

[109] Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)*,6(2):227.

[110] Jake Lever, Martin Krzywinski and Naomi Altman (2016) Logistic regression, *Nature Methods*, pages 541–542, https://doi.org/10.1038/nmeth.3904.

[111] Carina Mood (2010) Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It, *European Sociological Review*, Volume 26, Issue 1, Pages 67–82, https://doi.org/10.1093/esr/jcp006.

[112] Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016) Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1), 87-103, https://doi.org/10.1093/pan/mpv024.

[113] Mirjam J. Knol, Saskia Le Cessie, Ale Algra, Jan P. Vandenbroucke, Rolf H.H. Groenwold (2012) Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression, *CMAJ*, 184 (8) 895-899, https://doi.org/10.1503/cmaj.101715.

[114] Pijush Samui (2013) Multivariate Adaptive Regression Spline (Mars) for Prediction of Elastic Modulus of Jointed Rock Mass, *Geotechnical and Geological Engineering*, Volume 31, Issue 1, pp 249–253, https://doi.org/10.1007/s10706-012-9584-4.

[115] Pijush Samui, Sarat Das, Dookie Kim (2011) Uplift capacity of suction caisson in clay using multivariate adaptive regression spline,*Ocean Engineering, Volume 38, Issues 17–18*, Pages 2123-2127, https://doi.org/10.1016/j.oceaneng.2011.09.036.

[116] Kuhn Max; Johnson Kjell (2013) MARS regression, *Applied Predictive Modeling*, New York, NY: Springer New York.

[117] Statsoft (2019) Multivariate Adaptive Regression Splines (MARSplines), *Statsoft.com*, http://www.statsoft.com/Textbook/Multivariate-Adaptive-Regression-Splines.

[118] Pijush Samui,Pradeep Kurup (2012) Multivariate Adaptive Regression Spline and Least Square Support Vector Machine for Prediction of Undrained Shear Strength of Clay, *International Journal of Applied Metaheuristic Computing, 3(2)*, https://doi.org/10.4018/jamc.2012040103.

[119] Ariadna Montiel, Ruth Lazkoz, Irene Sendra, Celia Escamilla-Rivera, and Vincenzo Salzano (2014) Nonparametric reconstruction of the cosmic expansion with local regression smoothing and simulation extrapolation, *Physical Review D, 89, 043007*, https://doi.org/10.1103/PhysRevD.89.043007.

[120] Felix Biscarri, Inigo Monedero, Antonio Garcia, Juan Ignacio Guerrero, Carlos Leon (2017) Electricity clustering framework for automatic classification of customer loads, *Expert Systems with Applications, Volume 86*, Pages 54-63, https://doi.org/10.1016/j.eswa.2017.05.049.

[121] NIST/SEMATECH (2012) e-Handbook of Statistical Methods, *NIST/SEMATECH*, http://www.itl.nist.gov/div898/handbook.

[122] Kyoosik Kim (2019) Ridge Regression for Better Usage, *Towards data science*, https://towardsdatascience.com/ridge-regression-for-better-usage.

[123] C.B. Garcia, J. Garcia, M.M. Lopez Martín and R. Salmeron (2015) Collinearity: revisiting the variance inflation factor in ridge regression, *Journal of Applied Statistics, Volume 42 - Issue 3*, Pages 648-661, https://doi.org/10.1080/02664763.2014.980789.

[124] NCSS Statistical Software: Ridge regression, *NCSS*, https://ncss-wpengine.netdna-ssl.com.

[125] Patrick Breheny: Penalized regression methods, *University of Kentucky*, https://web.as.uky.edu/statistics/users/pbreheny.

[126] B M Golam Kibria, Shipra Banik (2016) Some Ridge Regression Estimators and Their Performances, *Journal of Modern Applied Statistical Methods, Volume 15, Issue 1 Article 12*.

[127] Prashant Gupta (2017) Regularization in Machine Learning, *Towards data science*, https://towardsdatascience.com/regularization-in-machine-learning.

[128] Gene H. Golub, Michael Heath and Grace Wahba (2012) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics, 21:2*, 215-223, https://doi.org/10.1080/00401706.1979.10489751.

[129] Charles K. Fisher, Austin Huang, and Collin M. Stultz (2010) Modeling Intrinsically Disordered Proteins with Bayesian Statistics, *Journal of the American Chemical Society, 132 (42)*, 14919-14927, https://doi.org/10.1021/ja105832g.

[130] Andre E. Punt and Ray Hilborn (2001) Strengths and weaknesses of Bayesian Approach, *Computerized Information Series*, Food and Organization of the United Nations.

[131] Hoffrage Ulrich, Krauss Stefan, Martignon Laura, Gigerenzer Gerd (2015) Natural frequencies improve Bayesian reasoning in simple and complex inference tasks, *Frontiers in Psychology, Volume 6*, pp.1473, https://doi.org/10.3389/fpsyg.2015.01473.

[132] Tina R. Patil, Mrs. S. S. Sherekar, Sant Gadgebaba (2013) Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal Of Computer Science And Applications Vol. 6, No.2*.

[133] Narayanan V., Arora I., Bhatia A. (2013) Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model, *Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL*, https://doi.org/10.1007/978-3-642-41278-3_24.

[134] S.L. Ting, W.H. Ip, Albert H.C. Tsang (2011) Is Naïve Bayes a Good Classifier for Document Classification?, *International Journal of Software Engineering and Its Applications Vol. 5, No. 3.*

[135] J. Zhang, C. Chen, Y. Xiang, W. Zhou and Y. Xiang (2013) Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions, *IEEE Transactions on Information Forensics and Security, vol. 8, no. 1*, pp. 5-15, https://doi.org/10.1109/TIFS.2012.2223675.

[136] Nayyar A. Zaidi, Jesus Cerquides, Mark J. Carman, Geoffrey I.Webb (2013) Alleviating Naive Bayes Attribute Independence Assumption by AttributeWeighting, *Journal of Machine Learning Research 14* , 1947-1988.

[137] Liangxiao Jiang, Chaoqun Li, Shasha Wang, Lungan Zhang (2016) Deep feature weighting for naive Bayes and its application to text classification,*Engineering Applications of Artificial Intelligence, Volume 52*, Pages 26-39, https://doi.org/10.1016/j.engappai.2016.02.002.

[138] Victor Roman (2019) Naive Bayes Algorithm: Intuition and Implementation in a Spam Detector, *Towards data science*, https://towardsdatascience.com/naive-bayes-intuition-and-implementation.

[139] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, Harry Zhang (2012) Improving Tree augmented Naive Bayes for class probability estimation,*Knowledge-Based Systems, Volume 26*, Pages 239-245, https://doi.org/10.1016/j.knosys.2011.08.010.

[140] Zhun Yu, Fariborz Haghighat, Benjamin C.M. Fung, Hiroshi Yoshino (2010) A decision tree method for building energy demand modeling, *Energy and Buildings, Volume 42, Issue 10*, Pages 1637-1646, https://doi.org/10.1016/j.enbuild.2010.04.006.

[141] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification, *ISPRS Journal of Photogrammetry and Remote Sensing, Volume 67*, Pages 93-104, https://doi.org/10.1016/j.isprsjprs.2011.11.002.

[142] Mahyat Shafapour Tehrany, Biswajeet Pradhan, Mustafa Neamah Jebur (2013) Spatial prediction of

flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS, *Journal of Hydrology, Volume 504*, Pages 69-79, https://doi.org/10.1016/j.jhydrol.2013.09.034.

[143] Jung Hwan Cho, Pradeep U. Kurup (2011) Decision tree approach for classification and dimensionality reduction of electronic nose data, *Sensors and Actuators B: Chemical, Volume 160, Issue 1*, Pages 542-548, https://doi.org/10.1016/j.snb.2011.08.027.

[144] Song YY, Lu Y (2015) Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry. 27(2)*, pp. 130 – 135.

[145] Bartosz Krawczyk, Michał Woźniak, Gerald Schaefer (2014) Cost-sensitive decision tree ensembles for effective imbalanced classification, *Applied Soft Computing, Volume 14, Part C*, Pages 554-562, https://doi.org/10.1016/j.asoc.2013.08.014.

[146] Tao Wang, Zhenxing Qin, Zhi Jin, Shichao Zhang (2010) Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning, *Journal of Systems and Software, Volume 83, Issue 7*, Pages 1137-1147, https://doi.org/10.1016/j.jss.2010.01.002.

[147] Rutvija Pandya, Jayati Pandya (2015) C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, *International Journal of Computer Applications Volume 117 – No. 16*, https://doi.org/10.5120/20639-3318.

[148] Shengyi Jiang, Guansong Pang, Meiling Wu, Limin Kuang (2012) An improved K-nearest-neighbor algorithm for text categorization, *Expert Systems with Applications, Volume 39, Issue 1*, Pages 1503-1509, https://doi.org/10.1016/j.eswa.2011.08.040.

[149] Sadegh Bafandeh Imandoust And Mohammad Bolandraftar (2013) Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background, *S B Imandoust et al. Int. Journal of Engineering Research and Applications, Vol. 3, Issue 5*, pp.605-610.

[150] X. Liang, X. Gou and Y. Liu (2012) Fingerprint-based location positoning using improved KNN, *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content, Beijing*, pp. 57-61.

[151] A. Thommandram, J. M. Eklund and C. McGregor (2013) Detection of apnoea from respiratory time series data using clinically recognizable features and kNN classification, *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka*, pp. 5013-5016, https://doi.org/10.1109/EMBC.2013.6610674.

[152] Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti (2013) Enhancing density-based clustering: Parameter reduction and outlier detection, *Information Systems, Volume 38, Issue 3*, Pages 317-330, https://doi.org/10.1016/j.is.2012.09.001.

[153] O. Kursun (2010) Spectral Clustering with Reverse Soft K-Nearest Neighbor Density Estimation, *The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona*, pp. 1-8.

[154] aporras (2018) 10 Reasons for loving Nearest Neighbors algorithm, *QuantDare*, https://quantdare.com/10-reasons\-for-nearest-neighbors-algorithm.

[155] Y. Mitani, Y. Hamamoto (2006) A local mean-based nonparametric classifier, *Pattern Recognition Letters, Volume 27, Issue 10*, Pages 1151-1159, https://doi.org/10.1016/j.patrec.2005.12.016.

[156] Gustavo E.A.P.A. Batista, Diego Furtado Silva (2009) How k-Nearest Neighbor Parameters Affect its Performance?, *Argentine symposium on artificial intelligence*.

[157] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, Xuelian Deng (2018) A novel kNN algorithm with data-driven k parameter computation, *Pattern Recognition Letters, Volume 109*, Pages 44-54, https://doi.org/10.1016/j.patrec.2017.09.036.

[158] Zhang, Shichao and Li, Xuelong and Zong, Ming and Zhu, Xiaofeng and Cheng, Debo (2017) Learning K for kNN Classification, *ACM Trans. Intell. Syst. Technol., Volume 8*, pp. 43:1–43:19, https://doi.org/10.1145/2990508.

[159] Abdulhamit Subasi (2013) Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, *Computers in Biology and Medicine, Volume 43, Issue 5*, Pages 576-586, https://doi.org/10.1016/j.compbiomed.2013.01.020.

[160] Fangjun Kuang, Weihong Xu, Siyang Zhang (2014) A novel hybrid KPCA and SVM with GA model for intrusion detection,*Applied Soft Computing, Volume 18*, Pages 178-184, https://doi.org/10.1016/j.asoc.2014.01.028.

[161] Roman M. Balabin, Ekaterina I. Lomakina (2011) Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data , *Analyst, issue:8*.

[162] S. Han, Cao Qubo and Han Meng (2012) Parameter selection in SVM with RBF kernel function, *World Automation Congress 2012, Puerto Vallarta, Mexico*, pp. 1-4.

[163] Weijun li, Zhenyu Liu (2011) A method of SVM with Normalization in Intrusion Detection, *Procedia Environmental Sciences, Volume 11, Part A*, Pages 256-262, https://doi.org/10.1016/j.proenv.2011.12.040.

[164] Adel Bolbol, Tao Cheng, Ioannis Tsapakis, James Haworth (2012) Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems, Volume 36, Issue 6*, Pages 526-537, https://doi.org/10.1016/j.compenvurbsys.2012.06.001.

[165] A. Bhardwaj, A. Gupta, P. Jain, A. Rani and J. Yadav (2015) Classification of human emotions from EEG signals using SVM and LDA Classifiers, *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), Noida*, pp. 180-185, https://doi.org/10.1109/SPIN.2015.7095376.

[166] De Giorgi, M.G., Campilongo, S., Ficarella, A., Congedo, P.M (2014) Comparison Between Wind Power Prediction Models Based on Wavelet Decomposition with Least-Squares Support Vector Machine (LS-SVM) and Artificial Neural Network (ANN) *Energies, 7*, 5251-5272, https://doi.org/10.3390/en7085251.

[167] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications, Volume 40, Issue 1* , Pages 200-210, https://doi.org/10.1016/j.eswa.2012.07.021.

[168] Trupti M. Kodinariya, Prashant R. Makwana (2013) Review on determining number of Cluster in K-Means Clustering, *International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6*.

[169] S. Na, L. Xumin and G. Yong (2010) Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, *Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan*, pp. 63-67.

[170] Soumi Ghosh, Sanjay Kumar Dubey (2013) Comparative analysis of k-means and fuzzy c-means algorithms, *International Journal of Advanced Computer Science and Applications, Vol. 4, No.4*, https://doi.org/10.14569/IJACSA.2013.040406.

[171] Taher Niknam, Babak Amiri (2010) An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis,*Applied Soft Computing, Volume 10, Issue 1* , pp. 183-197, https://doi.org/10.1016/j.asoc.2009.07.001.

[172] You Li, Kaiyong Zhao, Xiaowen Chu, Jiming Liu (2013) Speeding up k-Means algorithm by GPUs,*Journal of Computer and System Sciences, Volume 79, Issue 2*, pp. 216-229, https://doi.org/10.1016/j.jcss.2012.05.004.

[173] Steinley, D., and Brusco, M. J (2011) Choosing the number of clusters in K-means clustering, *Psychological Methods, 16(3)*, 285-297, https://doi.org/10.1037/a0023346.

[174] Renato Cordeiro de Amorim, Boris Mirkin (2012) Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, *Pattern Recognition, Volume 45, Issue 3* , pp. 1061-1075, https://doi.org/10.1016/j.patcog.2011.08.012.

[175] Carlos Ordonez, Edward Omiecinski (2004) Efficient Disk-Based K-Means Clustering for Relational Databases, *IEEE Transactions on Knowledge and Data Engineering, vol. 16*, pp. 909-921, https://doi.org/10.1109/TKDE.2004.25.

[176] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song (2015) Efficient agglomerative hierarchical clustering, *Expert Systems with Applications, Volume 42, Issue 5*, pp 2785-2797, https://doi.org/10.1016/j.eswa.2014.09.054.

[177] Dongkuan Xu, Yingjie Tian (2015) A Comprehensive Survey of Clustering Algorithms, *Annals of Data Science, Volume 2, Issue 2*, pp 165–193, https://doi.org/10.1007/s40745-015-0040-1.

[178] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, Saeed Ur Rehman (2014) Research on particle swarm optimization based clustering: A systematic review of literature and techniques, *Swarm and Evolutionary Computation, Volume 17*, Pages 1-13, https://doi.org/10.1016/j.swevo.2014.02.001.

[179] T. Nguyen and C. Kwoh (2015) Efficient agglomerative hierarchical clustering for biological sequence analysis, *TENCON 2015 - 2015 IEEE Region 10 Conference, Macao*, pp. 1-3.

[180] Guilherme Andrade, Gabriel Ramos, Daniel Madeira, Rafael Sachetto, Renato Ferreira, Leonardo Rocha (2013) G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering, *Procedia Computer Science, Volume 18*, Pages 369-378, https://doi.org/10.1016/j.procs.2013.05.200.

[181] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim, June Sup Lee (2014) DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce, *Information Systems, Volume 42*, Pages 15-35, https://doi.org/10.1016/j.is.2013.11.002.

[182] Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, Alfredo Pulvirenti (2013) Enhancing density-based clustering: Parameter reduction and outlier detection, *Information Systems, Volume 38, Issue 3*, Pages 317-330, https://doi.org/10.1016/j.is.2012.09.001.

[183] Sunita Jahirabadkar, Parag Kulkarni (2014) Algorithm to determine ϵ-distance parameter in density based clustering, *Expert Systems with Applications, Volume 41, Issue 6*, Pages 2939-2946, https://doi.org/10.1016/j.eswa.2013.10.025.

[184] A Amini, TY Wah (2011) Density micro-clustering algorithms on data streams: A review, *Proceedings of the International Multiconference of Engineers and Computer Scientists*.

[185] Dongwon Lee, Sung-Hyuk Park, Songchun Moon (2013) Utility-based association rule mining: A marketing solution for cross-selling, *Expert Systems with Applications, Volume 40, Issue 7*, Pages 2715-2725, https://doi.org/10.1016/j.eswa.2012.11.021.

[186] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen (2013) Association rule mining to detect factors which contribute to heart disease in males and females, *Expert Systems with Applications, Volume 40, Issue 4*, Pages 1086-1093, https://doi.org/10.1016/j.eswa.2012.08.028.

[187] R.J. Kuo, C.M. Chao, Y.T. Chiu (2011) Application of particle swarm optimization to association rule mining, *Applied Soft Computing, Volume 11, Issue 1*, Pages 326-336, https://doi.org/10.1016/j.asoc.2009.11.023.

[188] Zhang M., He C. (2010) Survey on Association Rules Mining Algorithms , *Advancing Computing, Communication, Control and Management. Lecture Notes in Electrical Engineering, vol 56. Springer, Berlin, Heidelberg.*, pp 111-118, https://doi.org/10.1007/978-3-642-05173-9_15.

[189] Andrew Tch: The mostly complete chart of Neural Networks, explained, *Towards data science*, https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464.

[190] H. Yu, T. Xie, S. Paszczynski and B. M. Wilamowski (2011) Advantages of Radial Basis Function Networks for Dynamic System Design, *IEEE Transactions on Industrial Electronics, vol. 58, no. 12*, pp. 5438-5450, https://doi.org/10.1109/TIE.2011.2164773.

[191] Mehdi Khashei, Mehdi Bijari (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, *Applied Soft Computing, Volume 11, Issue 2*, Pages 2664-2675, https://doi.org/10.1016/j.asoc.2010.10.015.

[192] Li-Hua Feng, Jia Lu (2010) The practical research on flood forecasting based on artificial neural networks, *Expert Systems with Applications, Volume 37, Issue 4*, Pages 2974-2977, https://doi.org/10.1016/j.eswa.2009.09.037.

[193] Daniel Westreich, Justin Lessler, Michele Jonsson Funk (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression, *Journal of Clinical Epidemiology, Volume 63, Issue 8*, Pages 826-833, https://doi.org/10.1016/j.jclinepi.2009.11.020.

[194] Ding, S., Su, C. and Yu (2011) An optimizing BP neural network algorithm based on genetic algorithm*Artificial Intelligence Review , Volume 36, Issue 2*, pp 153–162, https://doi.org/10.1007/s10462-011-9208-z.

[195] Hong-ze Li, Sen Guo, Chun-jie Li, Jing-qi Sun (2013) A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm, *Knowledge-Based Systems, Volume 37*, Pages 378-387, https://doi.org/10.1016/j.knosys.2012.08.015.

[196] SeyedAli Mirjalili, Siti Zaiton Mohd Hashim, Hossein Moradian Sardroudi (2012) Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm, *Applied Mathematics and Computation, Volume 218, Issue 22*, Pages 11125-11137, https://doi.org/10.1016/j.amc.2012.04.069.

[197] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov (2014) Dropout: a simple way to prevent neural networks from overfitting, *Journal of machine learning research*.

[198] Jonathan L. Ticknor (2013) A Bayesian regularized artificial neural network for stock market forecasting, *Expert Systems with Applications, Volume 40, Issue 14*, Pages 5501-5506, https://doi.org/10.1016/j.eswa.2013.04.013.

[199] Jurgen Schmidhuber (2015) Deep learning in neural networks: An overview, *Neural Networks, Volume 61*, Pages 85-117, https://doi.org/10.1016/j.neunet.2014.09.003.

[200] Chao Shang, Fan Yang, Dexian Huang, Wenxiang Lyu (2014) Data-driven soft sensor development based on deep learning technique, *Journal of Process Control, Volume 24, Issue 3*, Pages 223-233, https://doi.org/10.1016/j.jprocont.2014.01.012.

[201] Jie Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen and Chung-Ming Chen (2016)

Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans, *Scientific Reports volume 6*, https://doi.org/10.1038/srep24454.

[202] Ravid Shwartz-Ziv, Naftali Tishby (2017) Opening the Black Box of Deep Neural Networks via Information, *Machine learning*, https://arxiv.org/abs/1703.00810.

[203] Tanu Arya (2018) Drawbacks of Deep Learning, *Stanford Management Science and Engineering*, https://mse238blog.stanford.edu.

[204] P.K. Anooj (2012) Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, *Journal of King Saud University - Computer and Information Sciences, Volume 24, Issue 1*, Pages 27-40, https://doi.org/10.1016/j.jksuci.2011.09.002.

[205] Jose M. Alonso, Luis Magdalena (2011) Special issue on interpretable fuzzy systems, *Information Sciences, Volume 181, Issue 20*, Pages 4331-4339, https://doi.org/10.1016/j.ins.2011.07.001.

[206] Salma Elhag, Alberto Fernández, Abdullah Bawakid, Saleh Alshomrani, Francisco Herrera (2015) On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems,*Expert Systems with Applications, Volume 42, Issue 1*, Pages 193-202, https://doi.org/10.1016/j.eswa.2014.08.002.

[207] V. P. G. Jimenez, Y. Jabrane, A. G. Armada, B. Ait Es Said and A. Ait Ouahman (2011) High Power Amplifier Pre-Distorter Based on Neural-Fuzzy Systems for OFDM Signals, *IEEE Transactions on Broadcasting, vol. 57, no. 1*, pp. 149-158, https://doi.org/10.1109/TBC.2010.2088331.

[208] TE Alhanafy, F Zaghlool, A Saad, ED Moustafa (2010) Neuro-Fuzzy modeling scheme for the prediction of air pollution, *Journal of American Science, 6(12)*.

[209] Ilija Svalina, Vjekoslav Galzina, Roberto Lujic, Goran Simunovic (2013) An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices,*Expert Systems with Applications, Volume 40, Issue 15*, Pages 6055-6063, https://doi.org/10.1016/j.eswa.2013.05.029.

[210] B. Dennis, S. Muthukrishnan (2014) AGFS: Adaptive Genetic Fuzzy System for medical data classification, *Applied Soft Computing, Volume 25*, Pages 242-252, https://doi.org/10.1016/j.asoc.2014.09.032.

[211] P Amudha, S Karthik, S Sivakumari (2013) Classification techniques for intrusion detection an overview , *International Journal of Computer applications, Volume 76, No.16*, https://doi.org/10.5120/13334-0928.

[212] T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano (2011) Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 41, no. 3*, pp. 552-568, https://doi.org/10.1109/TSMCA.2010.2084081.

[213] Syarif I., Zaluska E., Prugel-Bennett A., Wills G (2012) Application of Bagging, Boosting and Stacking to Intrusion Detection, *International workshop on Machine Learning and Data Mining in Pattern Recognition. MLDM . Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg*, https://doi.org/10.1007/978-3-642-31537-4_46.

[214] GeeksforGeeks: Comparison b/w Bagging and Boosting in Data Mining, *https://www.geeksforgeeks.org/comparison-b-w-bagging-and-boosting-data-mining*.

[215] M. Paz Sesmero, Agapito I. Ledezma, Araceli Sanchis (2015) Generating ensembles of heterogeneous classifiers using Stacked Generalization, *Wires data mining and knowledge discovery*.

[216] Harshdeep Singh (2018) Understanding Gradient Boosting Machines, *https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab*.

[217] Ferreira A.J., Figueiredo M.A.T. : Boosting Algorithms: A Review of Methods, Theory, and Applications, *Ensemble Machine Learning. Springer, Boston, MA*, pp 35-85, https://doi.org/10.1007/978-1-4419-9326-7_2.

[218] Wilson, Andrew G and Gilboa, Elad and Nehorai, Arye and Cunningham, John P (2014) Fast Kernel Learning for Multidimensional Pattern Extrapolation, *Advances in Neural Information Processing Systems 27*, pp.3626- 3634.

[219] Wang S () CyberGIS and spatial data science, *GeoJournal, Volume 81, Issue 6*, pp.965–968, https://doi.org/10.1007/s10708-016-9740-0.

[220] William Q. Meeker and Yili Hong (2014) Reliability Meets Big Data: Opportunities and Challenges, *Quality Engineering, 26:1*, 102-116, https://doi.org/10.1080/08982112.2014.846119.

[221] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais and Prabhat (2019) Deep learning and process understanding for data-driven Earth system science, *Nature, volume 566*, pages195–204, https://doi.org/10.1038/s41586-019-0912-1.

[222] Peter V. Coveney, Edward R. Dougherty and Roger R. Highfield (2016) Big data need big theory too, *Philosophical transactions of the Royal Society A Mathematical, Physical and Engineering sciences*.

[223] Xiang Liu, Ziyang Tang, Huyunting Huang, Tonglin Zhang and Baijian Yang (2019) Multiple Learning for Regression in big data, *CoRR*.

[224] Ping Ma, Xiaoxiao Sun (2014) Leveraging for big data regression, *Wires Computational Statistics*.

[225] S Jun, SJ Lee, JB Ryu (2015) A divided regression analysis for big data, *International Journal of software engineering and its applications*.

[226] HaiYing Wang, Min Yang and John Stufken (2019) Information-Based Optimal Subdata Selection for Big Data Linear Regression, *Journal of the American Statistical Association, volume 114, number 525*, pp. 393-405, https://doi.org/10.1080/01621459.2017.1408468.

[227] Yang, Hang and Fong, Simon (2012) Incrementally Optimized Decision Tree for Noisy Big Data, *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*,pp.36-44.

[228] W Dai, W Ji (2014) A mapreduce implementation of C4. 5 decision tree algorithm, *International journal of database theory and application, Vol 7, No. 1*, pp.49-60, https://doi.org/10.14257/ijdta.2014.7.1.05.

[229] J. Chen et al. (2017) A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment, *IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4*, pp. 919-933, https://doi.org/10.1109/TPDS.2016.2603511.

[230] Ke, Guolin and Meng, Qi and Finley, Thomas and Wang, Taifeng and Chen, Wei and Ma, Weidong and Ye, Qiwei and Liu, Tie-Yan (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems 30*, pp. 3146-3154.

[231] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, Shichao Zhang (2016) Efficient kNN classification algorithm for big data, *Neurocomputing, Volume 195*, Pages 143-148, https://doi.org/10.1016/j.neucom.2015.08.112.

[232] Jesus Maillo, Sergio Ramírez, Isaac Triguero, Francisco Herrera (2017) kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data,*Knowledge-Based Systems, Volume 117*, Pages 3-15, https://doi.org/10.1016/j.knosys.2016.06.012.

[233] J. Maillo, I. Triguero and F. Herrera (2015) A MapReduce-Based k-Nearest Neighbor Approach for Big Data Classification, *IEEE Trustcom/BigDataSE/ISPA, Helsinki*, pp. 167-172.

[234] X Yan, Z Wang, D Zeng, C Hu and H Yao (2014) Design and analysis of parallel MapReduce based KNN-join algorithm for big data classification, *TELKOMNIKA Indonesian Journal of ELectrical Engineering, Vol 12, No 11* , pp.7927-7934, https://doi.org/10.11591/telkomnika.v12i11.6357.

[235] G. Song, J. Rochas, L. E. Beze, F. Huet and F. Magoules (2016) K Nearest Neighbour Joins for Big Data on MapReduce: A Theoretical and Experimental Analysis, *IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9*, pp. 2376-2392, https://doi.org/10.1109/TKDE.2016.2562627.

[236] Katkar, V. D., and Kulkarni, S. V. (2013) A novel parallel implementation of Naive Bayesian classifier for Big Data, *International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*

[237] B. Chandra, Manish Gupta (2011) Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data, *Expert Systems with Applications,Volume 38, Issue 3*,Pages 1293-1298, https://doi.org/10.1016/j.eswa.2010.06.076.

[238] Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017) Machine learning on big data: Opportunities and challenges, *Neurocomputing, 237*, pp.350–361, https://doi.org/10.1016/j.neucom.2017.01.026.

[239] Sergio Ramirez Gallego ,Salvador Garcia, Hector Mourino-Talin , David Martinez-Rego , Veronica Bolon-Canedo ,Amparo Alonso-Betanzos ,Jose Manuel Benitez ,Francisco Herrera (2015) Data discretization: taxonomy and big data challenge, *Advanced Review – Wires Data mining and knowledge discovery*, https://doi.org/10.1002/widm.1173.

[240] Rebentrost, Patrick and Mohseni, Masoud and Lloyd, Seth (2014) Quantum Support Vector Machine for Big Data Classification, *Phys. Rev. Lett. – Volume 113, Issue 13, pp. 130503*, https://doi.org/10.1103/PhysRevLett.113.130503.

[241] Anushree Priyadarshini, SonaliAgarwal (2015) A Map Reduce based Support Vector Machine for Big Data Classification, *International Journal of Database Theory and Application , Vol.8 No.5* ,pp.77-98, https://doi.org/10.14257/ijdta.2015.8.5.07 .

[242] D. Singh, D. Roy and C. K. Mohan : DiP-SVM (2017) Distribution Preserving Kernel Support Vector Machine for Big Data, *IEEE Transactions on Big Data, vol. 3, no. 1*, pp. 79-90, https://doi.org/10.1109/TBDATA.2016.2646700.

[243] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., and Plaza, A (2015) On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(10)*, 4634–4646, https://doi.org/10.1109/JSTARS.2015.2458855.

[244] Y. Liu and J. Du (2015) Parameter Optimization of the SVM for Big Data, *8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou*, pp. 341-344.

[245] Xiao Cai, Feiping Nie, Heng Huang (2013) Multi-View K-Means Clustering on Big Data, *Twenty-Third International Joint Conference on Artificial Intelligence, Web and Knowledge-Based Information Systems*.

[246] Cui, X., Zhu, P., Yang, X., Li, K., and Ji, C. (2014) Optimized big data K-means clustering using MapReduce, *The Journal of Supercomputing, 70(3)*, pp.1249–1259, https://doi.org/10.1007/s11227-014-1225-7.

[247] N. Akthar, M. V. Ahamad and S. Khan (2015) Clustering on Big Data Using Hadoop MapReduce, *International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur*, pp. 789-795.

[248] Rathore, Punit and Kumar, Dheeraj and C. Bezdek, James and Rajasegarar, Sutharshan and Palaniswami, Marimuthu (2018) A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data , *IEEE Transactions on Knowledge and Data Engineering* PP. 10.1109.

[249] T. C. Havens, J. C. Bezdek and M. Palaniswami (2013) Scalable single linkage hierarchical clustering for big data, *IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, VIC*, pp. 396-401.

[250] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song (2015) Efficient agglomerative hierarchical clustering, *Expert Systems with Applications,Volume 42, Issue 5*,Pages 2785-2797, https://doi.org/10.1016/j.eswa.2014.09.054.

[251] Yunpeng Cai, Yijun Sun, ESPRIT-Tree (2011) Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time, *Nucleic Acids Research, Volume 39, Issue 14*, Page e95, https://doi.org/10.1093/nar/gkr349.

[252] Embrechts, M and Gatti, Christopher and Linton, Jonathan and Roysam, Badrinath (2013) Hierarchical Clustering for Large Data Sets , *Advances in Intelligent Signal Processing and Data Mining: Theory and Applications*, pp.197-233, https://doi.org/10.1007/978-3-642-28696-4_8.

[253] Yanwei Yu, Jindong Zhao,Xiaodong Wang, Qin Wang , Yonggang Zhang (2015) Cludoop: An Efficient Distributed Density-Based Clustering for Big Data Using Hadoop, *International Journal of Distributed Sensor Networks*.

[254] Amini, A., Wah, T.Y. and Saboohi H. J (2014) On Density-Based Data Streams Clustering Algorithms: A Survey, *Journal of Computer Science and Technology - Volume 29, Issue 1*, pp 116–141, https://doi.org/10.1007/s11390-014-1416-y.

[255] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim, June Sup Lee (2014) DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce, *Information Systems, Volume 42*,Pages 15-35, https://doi.org/10.1016/j.is.2013.11.002.

[256] Deng, Z., Hu, Y., Zhu, M. et al (2015) A scalable and fast OPTICS for clustering trajectory big data, *Cluster Computing 18: 549*, https://doi.org/10.1007/s10586-014-0413-9.

[257] Imran Khan, Joshua Z. Huang, Kamen Ivanov (2016) Incremental density-based ensemble clustering over evolving data streams,*Neurocomputing, Volume 191*,Pages 34-43, https://doi.org/10.1016/j.neucom.2016.01.009.

[258] Wang, Yu and Li, Boxun and Luo, Rong and Chen, Yiran and Xu, Ningyi and Yang, Huazhong (2014) Energy efficient neural networks for big data analytics, *Proceedings of the Conference on Design, Automation and Test in Europe*.

[259] Cao J, Cui H, Shi H, Jiao L : Big Data (2016) A Parallel Particle Swarm Optimization Back Propagation Neural Network Algorithm Based on MapReduce,*PLoS ONE 11(6): e0157551*, https://doi.org/10.1371/journal.pone.0157551.

[260] Chiroma, Haruna, Ali Abdullahi, Usman, Abdulhamid, Shafi i, Abdulsalam AlArood, Ala and Gabralla, Lubna and Rana, Nadim and Shuib, Liyana and Hashem, Ibrahim and Dada, Emmanuel and Abubakar, Adamu and Zeki, Akram and Herawan, Tutut (2018) Progress on Artificial Neural Networks for Big Data Analytics: A Survey, *IEEE Access. PP. 1-1*.

[261] Zhang, Y Guo, Q Wang, J : Big data analysis using neural networks 49. 9-18, *10.15961/j.jsuese.2017.01.002*.

[262] Hai Wang, Zeshui Xu, Witold Pedrycz (2017) An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities, *Knowledge-Based Systems, Volume 118*, Pages 15-30, https://doi.org/10.1016/j.knosys.2016.11.008.

[263] Alberto Fernandez and Cristobal Jose Carmona and Maria Jose del Jesus and Francisco Herrera (2016) A View on Fuzzy Systems for Big Data: Progress and Opportunities, *International Journal of Computational Intelligence Systems, Volume 9*, pp.69-80, https://doi.org/10.1080/18756891.2016.1180820.

[264] X. Chen and X. Lin (2014) Big Data Deep Learning: Challenges and Perspectives, *IEEE Access, vol. 2*, pp. 514-525, https://doi.org/10.1109/ACCESS.2014.2325029.

[265] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic (2015) Deep learning applications and challenges in big data analytics, *Journal of Big Data, 2:1*, https://doi.org/10.1186/s40537-014-0007-7.

[266] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T Dudley (2018) Deep learning for healthcare: review, opportunities and challenges, *Briefings in Bioinformatics, Volume 19, Issue 6*, Pages 1236–1246, https://doi.org/10.1093/bib/bbx044.

[267] Bartosz Krawczyk, Leandro L. Minku, Joao Gama, Jerzy Stefanowski, Michal Wozniak (2017) Ensemble learning for data stream analysis: A survey, *Information Fusion,Volume 37*,Pages 132-156, https://doi.org/10.1016/j.inffus.2017.02.004.

[268] Shan Huang, Botao Wang, Junhao Qiu, Jitao Yao, Guoren Wang, Ge Yu (2016) Parallel ensemble of online sequential extreme learning machine based on MapReduce, *Neurocomputing,Volume 174, Part A*,Pages 352-367, https://doi.org/10.1016/j.neucom.2015.04.105.

[269] Huang X., Ye Y., Zhang H. (2016) Extending Kmeans-Type Algorithms by Integrating Intra-cluster Compactness and Inter-cluster Separation, *Unsupervised Learning Algorithms. Springer*, pp 343-384, https://doi.org/10.1007/978-3-319-24211-8_13.

[270] N. B. Nikhare and P. S. Prasad (2018) A review on inter-cluster and intra-cluster similarity using bisected fuzzy C-mean technique via outward statistical testing, *2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore*, pp. 215-217.

[271] Frederic Godin, Jonas Degrave, Joni Dambre, Wesley De Neve (2018) Dual Rectified Linear Units (DReLUs): A replacement for tanh activation functions in Quasi-Recurrent Neural Networks, *Pattern Recognition Letters, Volume 116*, Pages 8-14, https://doi.org/10.1016/j.patrec.2018.09.006.

[272] Yu Wang (2017) A new concept using LSTM Neural Networks for dynamic system identification, *American Control Conference (ACC), Seattle, WA*, pp. 5324-5329, https://doi.org/10.23919/ACC.2017.7963782.

[273] Igor M. Coelho, Vitor N. Coelho, Eduardo J. da S. Luz, Luiz S. Ochi, Frederico G. Guimaraes, Eyder Rios (2017) A GPU deep learning metaheuristic based model for time series forecasting, Applied Energy, Volume 201, Pages 412-418, https://doi.org/10.1016/j.apenergy.2017.01.003.

[274] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014) Generative Adversarial Networks (PDF). Proceedings of the International Conference on Neural Information Processing Systems (NIPS). pp. 2672–2680.

[275] Michael A Farnum, Lalit Mohanty, Mathangi Ashok, Paul Konstant, Joseph Ciervo, Victor S Lobanov, Dimitris K Agrafiotis (2019) A dimensional warehouse for integrating operational data from clinical trials, *Database, Volume 2019*, https://doi.org/10.1093/database/baz039.

[276] Khan S.I., Hoque A.S.M.L (2015) Towards Development of National Health Data Warehouse for Knowledge Discovery, *Part of Advances in Intelligent Systems and Computing, vol 385. Springer, Cham*.

[277] SO Salinas, ACN Lemus (2019) Data warehouse and big data integration, *International Journal of Computer Science and Information Technology, Vol 9, No.2*.

[278] Hai, Rihan and Geisler, Sandra and Quix, Christoph (2016) Constance: An Intelligent Data Lake System, *Proceedings of the 2016 International Conference on Management of Data*.