Chapter – 2

Literature Review

This chapter contains an in-depth study of Data mining techniques proposed by the research community and explore the capabilities that reside in the integrated approach of data mining techniques. The state-of-the-art data mining techniques are reviewed in the first section. The tools used for Data mining are presented in the second section. The techniques and methods used to improve the performance of any traditional classifier are represented in the Third section. Discussion on the techniques used for handling imbalanced data by researchers has been dealt with in the fourth section and in the last section methods of feature selection are surveyed.

2.1 Data Mining Techniques

Mduma, Kalegele, and Machuve (2019) reported the highlights of open challenges for future research directions. A survey on machine learning techniques used in the fight against dropouts is presented. The survey conducted aids in the provision of a niche for students, researchers, and developers who aspire to apply the techniques.

The aim is to identify and use them to predict default risks, avoid possible payment difficulties, and reduce potential problems in extending credit data mining classification algorithms. This study includes demographic and socioeconomic characteristics of individuals were obtained from the Turkish Statistical Institute 2015 survey. Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression were the six classification algorithms applied to the dataset using WEKA 3.9 data mining software. These algorithms were compared considering the root mean error squares, accuracy, F-measure receiver

operating characteristic area, precision, and recall statistical criteria. To the real dataset, the best algorithm logistic regression was obtained and applied to determine the attributes causing the default risk by using odds ratios. The socioeconomic and demographic characteristics of the individuals were examined and based on the odds ratio values, the individuals' results and characteristics of which were likely to default, were reached as reported by Begum Çıgsar (2019).

The analysis of the performance of various classifiers like K-nearest neighbor, Naïve Bayes, generalized linear model, Gradient boosted trees, deep learning with H2O. Four synthesized datasets were used to check the classifiers. The Rapid miner tool was used to carry out this experiment. Deep learning with H2O outperforms the other classifiers in most of the cases that were observed in the results. These results were discussed (Arunadevi, 2018).

Raman et al. (2017) discuss the measures and the performance analysis of the classification algorithms using the WEKA tool. Chen et al. (2015) reviewed on systematic work on data mining in technical and knowledge view, and application view, including classification, clustering, association analysis, time series analysis, and outlier analysis.

IoT and big data are discussed comprehensively; survey the new technologies to mine big data in IoT, over viewing the challenges in the era of big data, and a new proposal of a big data mining system architecture for IoT. The contribution of the present paper includes 3 parts: the first part is a proposal to a novel way of reviewing data mining in the technique view, knowledge view, and application view; the second part is a discussion on the new characteristics of big data & its analysis along with the challenges. Another important contribution proposed is a big data mining system (Chen, 2015).

The basic definitions of clustering and the typical procedure, lists the commonly used distance (dissimilarity) functions, similarity functions, and evaluation

indicators that lay the foundation of clustering, and analyzes the clustering algorithms from two perspectives, the traditional ones that contain 9 categories including 26 algorithms and the modern ones that contain 10 categories including 45 algorithms The main purpose of the paper is to introduce the basic and core idea of each commonly used the advantages and disadvantages of analyzing clustering algorithm, specify the source of each one, and each one was reported in (Anna,2015).

Zehra et al. (2014) have worked on the importance of data preprocessing in data mining .the results show a significant improvement in the accuracy of preprocessed data over not preprocessed data.

Jain (2010) provided a brief overview of clustering, summarize well-known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and ensemble clustering, including semi-supervised clustering, and simultaneous feature selection during data clustering, point out some of the developing and useful research directions, and large scale data clustering.

Mariscal et al. (2010) described the most used industrial and academic projects which are cited in the scientific literature. They have also discussed data mining and knowledge discovery methodologies and process models, providing an overview of its knowledge discovery, evolution, history along with data mining and setting with the state- of- the –art topic. For an individual approach, a brief description of the proposed knowledge discovery in databases (KDD) process, discussing special features, outstanding advantages, and disadvantages of every approach. Apart from that, a global comparison of all presented data mining approaches is provided, focusing mainly on the different steps and tasks of every approach that interprets the whole KDD process.

Kotsiantis (2007) studied various supervised machine learning classification techniques. He emphasizes the major theoretical issues, guiding in interesting research directions and suggesting possible bias combinations that have yet to be explored.

Rui (2005) surveyed clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and their applications set some benchmark in data sets, like the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several related topics viz., cluster validation, and proximity measures were also discussed.

Sherry (2004) studied the impacts of data mining by reviewing existing applications, including personalized environments, electronic commerce, and search engines. How data mining can enhance the functions of the three types of applications, were discussed. This paper gives an overview of the state of the art research associated with these applications. With the limitations of current work and it raises several directions for future research.

Information retrieval is the most important aspect of data mining; the techniques being classification, association rules, and clustering. An attempt has been made here to discuss all these techniques (Rao, 2003).

Lee et al. (2001) discussed the challenges of DM and major DM techniques such as statistics, decision tree approach, genetic algorithm, artificial intelligence, and visualization.

Elder (2001) worked on major filtering approaches such as texture feature extraction and performed a comparative study. The filtering effect is highlighted, keeping the local energy function identical for most of the cases.

Dietterich (2000) studied the ensemble method with Bayesian averaging. This study emphasizes and explains why ensembles can often perform better than any single classifier. Some comparative studies on ensemble methods are reviewed, and some new experiments are presented to uncover the details. The work done had provided some experimental results in elucidating one of the reasons for AdaBoost's performance is good. One concern is that the interaction between AdaBoost and the underlying learning algorithm properties. Most of the learning algorithms have combined AdaBoost with the algorithms of a global character.

2.2 Data Mining Tools

literature survey in an analysis of the huge volume of healthcare-related data was carried out by Sharma et al. (2017) in prediction and diagnosis of lifestyle diseases the objective of this work is to study the different data mining techniques which can be employed in automated lifestyle diseases prediction systems. Various techniques and data mining classifiers are defined in the proposed work that had emerged in recent years for efficient and effective heart disease and types II diabetes diagnosis. This work provides a comparative account on a study of many data mining tools and data mining techniques like Decision tree, SVM, Naïve Bayes, Neural Network over disease prediction system. The experimental results show that many of the rules and techniques help in the best prediction of disease.

This study aims to improve information retrieval activities to a higher level. Different methods for information retrieval have been studied and discussed. That includes the use of the Fuzzy Ontology Generation framework (FOGA) framework and Formal Concept Analysis (FCA) based clustering and keyword matching approach. For retrieval of data from search engines, the Hidden Markov Model has been used intelligently and efficiently for correct identification and retrieval from the database. For the detection of community and conversion

of large community graphs to sub-community graphs for its better study and usage, The classification algorithm has been used (Verma, 2016).

Sethunya et al. (2016) researched on how data mining has successfully yielded numerous tools, algorithms, methods, and approaches for handling large amounts of data for various purposeful uses and problem-solving. Data mining is an integral part of many applications such as data warehousing, predictive analytics, business intelligence, and bioinformatics and decision support systems. Data mining's objective is to effectively handle large scale data, extract actionable patterns, and gain insightful knowledge. It is part of knowledge discovery in databases (KDD) process; the present paper gives an overview of various algorithms necessary for handling large data sets. To handle big data the algorithms used to define the various structures and methods implemented. The strengths and limitations of these algorithms have been reviewed and discussed. The paper will be a quick guide or an eye-opener to the data mining researchers on which algorithm(s) to select and apply in solving the problems they will be investigating.

A Comparison Study between Data Mining Tools over some Classification Methods (Abdullah, 2011), helped us gain knowledge about the implementation of classification algorithms in these tools and gave a fair idea on how to work on the study proposed. The research is very useful in building a platform for the comparative study undertaken.

Rathee and Mathur (2013) studied the education database, that contained hidden knowledge for improving student's performance. The comparative study of decision tree algorithms like ID3, C4.5, and CART, and as a result, C4.5 is more accurate. The predictions obtained from the algorithms help the teacher to segregate poor students and improve their performance.

Arora (2012) used two classification algorithms J48 (which is a java implementation of C4.5 algorithm) and multilayer perceptron alias MLP (which is a

modification of the standard linear perceptron) of the Weka interface. Apparently, used for testing several datasets. The performance of J48 and Multilayer Perceptron has been analyzed to choose the better algorithm based on the conditions of the datasets. The datasets were chosen from the UCI Machine Learning Repository. Algorithm J48 is based on C4.5 decision-based learning and algorithm Multilayer Perceptron uses the multilayer feed-forward neural network approach for classification of datasets. Comparatively, in the performance of both algorithms, Multilayer Perceptron was found to be a better algorithm in most of the cases.

Xindong et al. (2007) KNN classification is easy to understand and easy to implement classification techniques. Being simple, it is performing well in many situations.

Quinlan (1993) reported Iterative Dichotomiser or ID3 is a simple decision tree learning algorithm. C4.5 algorithm is an improved version of ID3; the gain ratio here is as splitting criteria. The difference between ID3 and C4.5 algorithm is that ID3 uses binary splits, whereas the C4.5 algorithm uses multi-way splits.

Larose (2005) discussed the KNN (K-Nearest Neighbor) algorithm, introduced by the Nearest Neighbor algorithm which is designed to find the nearest point of the observed object. The K-nearest points are found by the KNN algorithm.

There are a lot of improvements in the traditional KNN algorithm, such as the Wavelet-Based K-Nearest Neighbor Partial Distance Search (WKPDS) algorithm Equal Average Nearest Neighbor Search (ENNS) algorithm (Hwang, 2004).

Bernhard Schölkopf (1998) in an introductory overview, points out a particular advantage of SVMs over other learning algorithms is that it can be analyzed theoretically using concepts from computational learning theory, with good

performance is achieved when applied the same time to real problems. Examples of these real-world applications are provided by Sue Dumais (1998), mentioned text-categorization problem, that yields the best results to date on the Reuters collection, and Edgar Osuna (1998), who presents strong results on application to face detection. John Platt gives us a practical guide and a new technique for implementing the algorithm efficiently.

The paper reviews applications of data mining in manufacturing engineering viz., production processes, operations, fault detection, maintenance, decision support, and product quality improvement. Customer relationship management, standardization, and information integration aspects are also briefly discussed. The present review is focused on demonstrating the relevancy of data mining to the manufacturing industry, rather than discussing the data mining domain in general. The volume of general data mining literature makes it difficult to gain a precise view of a target area such as manufacturing engineering, which has its own particular needs and requirements for mining applications. This review reveals progressive applications in addition to existing gaps and fewer cons (Hearst et al., 1998).

The studies on Heart disease diagnosis that has been done by various data mining methods. Heart disease diagnosis by classification methods such as J48, REPTREE, Naïve Bayes, Bayes Net, and Simple CART (Machete, 2014). Heart Attack Prediction System through the K-means clustering algorithm on the pre-processed data and the recurrent patterns are mined with the MAFIA algorithm (Patil, 2009). Heart Disease Diagnosis Using Fuzzy Logic Approach (Anooj, 2012). It automatically retrieves the knowledge from the patient clinical data. Heart Disease Prediction Using Association Rule Manikandan et al. (2013) extract the item set relations by using association rule. Heart Disease Prediction System using Hybrid System Chitra et.al. (2013) Present Hybrid Intelligent techniques for the prediction of heart disease. Feature reduction improves

classification accuracy. Hybrid Model was used in the Prediction of Type II Diabetes.

Jayaram et al. (2012) develop a hybrid model for classifying Pima Indian Diabetic Database (PIDD). The model has two stages: the first stage comprised of the K-means clustering was used to identify and eliminated incorrectly classified instances. The second stage is a fine-tuned classification using a Decision tree C4.5 by taking them a correctly clustered instance of the first stage. Experimental results signify that cascaded K-means clustering and the rules generated by cascaded C4.5 tree with categorical data are easy to interpret as compared to rules generated with C4.5 alone with continuous data. The cascaded model with categorical data obtained Type II Diabetes Prediction Using Classification Han et al. used Rapid Miner for the classification of diabetes data analysis and diabetes prediction model. ID3 algorithm was used for prediction with 72% and 80% accuracy for which a Decision tree was generated respectively (Han,2008).

Type- II Diabetes Prediction Using Rough Sets on the PIMA for the first time. At first, the data was pre-processed and discredited by making intervals of data. Use of the equal frequency binning criteria for intervals and then the creation of reducts by using the Johnson reducer algorithm and classification using the batch classifier with the standard/tuned voting method (RSES). For each of the 10 randomizations of the PIDD training sets the rules were constructed from above Type II Diabetes Prediction.

Vijayalakshmi, (2012) developed a clustering algorithm that is used for predicting diabetes based on graph b-coloring technique. They implement and perform experiments by comparing their approach with K-NN classification and K-means clustering. The results showed that the clustering based on graph coloring is much better than other clustering approaches in terms of accuracy and purity. A real representation of clusters by dominant objects, assures the inter-cluster disparity in evaluating the quality of clusters and partitioning

Quinlan (1993) Iterative Dichotomiser-3 or ID3 is a simple decision tree learning algorithm C4.5 algorithm is an improved version of ID3; gain ratio is used as splitting criteria. The variation between ID3 and C4.5 algorithm is that binary splits are used in ID3, whereas the C4.5 algorithm uses multi-way splits.

Jason et al. (2003) proposed simple, heuristic solutions to some of the problems in Naive Bayes classifiers, addresses the systemic issues as well as problems that arise because the text is not actually generated according to a multinomial model. We find that our simple corrections result in a fast algorithm that is competitive with state-of-the-art text classification algorithms such as the Support Vector Machine We has described several techniques. These modifications better align Naive Bayes with the realities of bag-of-words textual data and, as we have shown empirically, significantly improve its performance on several data sets. The newer version of Naive Bayes is fast and is easy-to-implement.

Table - 2. 1 presents a Comparative analysis of six popular Data mining Tools used by the data mining community. It gives the information about the date of release of the tool, the name of the operating system on which it can be operated, the name of the language used to develop the tool, resource to download the tool, the application areas of the tool, acceptable data types and basic features.

Table - 2. 1 Comparative analysis of Data mining Tools

No	Name	Release date	Operating System	Language	Resource	Type	Data Source	Features
1	RAPID MINER	2006	Cross-platform	Language - Independent	www.rapidminer.com	Statistical & predictive analysis, Data mining,	ARFF ,MySQL, Access, CSV. ODBC, JDBC	More than 20 functions for handling and analysis of data. •various kind of File operators • A macro viewer to view the values during execution. • Interactive GUI
2	KNIME	2004	Linux ,OS X, Windows	Java	www.knime.org	Enterprise Reporting, Business Intelligence, Data mining	ARFF ,MySQL, Access, CSV, ODBC, JDBC	Scalability, inbuilt user interface, High extensibility • well-defined API for plug-in extensions • sophisticated data handling, Data visualization • Import/Export of workflows, Parallel execution on multi-core systems
3	ORANGE	2009	Cross-platform	Python C++, C	www.orange.biolab.si	Machine learning & Data mining,	SQL	Visual Programming, •Interaction And Data Analytics • Rich toolbox, a Scripting interface • Extendable Documentation
4	KEEL	2004	Cross-platform	Java	www.sci2s.ugr.es/keel	Machine Learning.	DAT, csv	Classification, Cluster, Visualization, Regression, Association, evolutionary learning , user-friendly

								graphical interface,
5	WEKA	1993	Cross platform	Java	www.cs.waikato.ac.nz/~ml/weka	Machine Learning.	ARFF, CSV, Excel	49 tools for data preprocessing 76 classification & regression algorithms, 8 clustering algorithms, 3 algorithms for finding association rules. 15 attribute/subset evaluators 10 search algorithms for feature selection. 3 graphical user interfaces • The Explorer • The Experimenter • The Knowledge Flow
6	R	1997	Cross-platform	C, Fortran, and R	www.r-project.org	Statistical Computing	ARFF, CSV, Excel	Data Exploration, Clustering, Time Series Analysis, Text Mining, Outlier detection, Social Network Analysis, Visualization of geospatial data, Parallel Computing, Graphics, Web Application & Big data

2.3 Improve the Performance of Classification Algorithm

Zeng et al. (2003) presented a clustering-based classification (CBC) approach when sufficient labeled data is not present as most of the traditional text classification algorithm's accuracy degrades drastically on an insufficient amount of labeled data. They have proposed an approach that is more effective with small training data and at the same time easier to achieve high performance especially when the labeled data is sparse. They have cascaded K-means and TSVM clustering and classification algorithms and applied them on 20- Newsgroups, Reuters-21578, and Open Directory Project (OPD) WebPages. The approach initially makes clusters of labeled and unlabeled data and according to the cluster, results expand the labeled set because of more the labeled data greater the accuracy. They have used P=100% for his experiments (means all unlabeled data are labeled by the clustering results). The result reveals that the performance of the CBC is superior over the existing method when labeled data size is very small in all other cases the algorithm is not adaptable.

Karegowda (2012) they have proposed a 3 stage model for diabetic Patient datasets. At the first stage, the K-means clustering algorithm was used to identify and eliminate incorrectly classified instances. In the next stage, they have extracted relevant features by cascading the Genetic algorithm and correlation-based feature selection (CFS) algorithm. And at the final stage classification is done using K-nearest neighbor (KNN) on the filtered datasets of stage -1 and stage-2. Experiments were carried out for different values of K ranging from 1 to 15 with GA-CFS feature selection and the accuracy of 96.67% with K=5 is reported.

Asha (2010)This paper proposes a combined approach of clustering and classification for the detection of tuberculosis (TB) patients'-means clustering algorithm with Naïve Bayes, C4.5 decision tree, Support Vector Machine, AdaBoost and Random Forest tree are combined to improve the accuracy. Initially,

the data is clustered in two sets, and respective classes are assigned to them. Then various classification algorithms are trained with these clusters based on the K fold cross-validation method. The best-obtained accuracy is 98.7% with Support Vector Machine is reported.

Shekhar (2007) they have proposed a cascaded method for classifying anomalous activities in a computer network, a machine mass beam an active electronic circuit. Initially, training instances are portioned into K clusters now on each cluster the decision boundaries are further refined by learning the subgroup within the cluster. At the final stage to obtain a final decision on classification two rules are deployed to combine K-means and ID3 1) The Nearest Neighbor rule 2) the nearest consensus rule then they compared the proposed cascaded approach with the individual K-means and ID3 method over six significant parameters on 3 datasets. Results show that the proposed model outperforms in terms of all 6 performance measures over 3 datasets.

The paper used Term frequency- Inverse document frequency (Tf-IDF) to weigh the words then group the datasets into clusters by the K-means clustering algorithm. The cluster centers are used as the new training sample for the KNN classification algorithm (PutuWira, 2012).

Mohammad (2009) the paper presents an integrated approach of subspace clustering and K nearest neighbor for text classification. The authors proposed an innovation by applying the impurity component for measuring dispersions and chi-square statistics for dimensions of a cluster. The experiments are conducted on NSF abstract datasets and 20 Newsgroup datasets .the comparisons are done on ROC (Receiver Operating Characteristics curve). The proposed model achieved an AUC (Area under the ROC curve) value of .92 while other methods recorded highest as .89 for NSF abstract datasets and for 20 newsgroup datasets AUC value is .813 while others can achieve .77 as highest.

López (2012) presented an approach to predict the performance of the students based on the usages of Forum data. The objective of their study is to determine whether participation in the courses forum can be a good predictor to evaluate the performance in the examination and how well the proposed integrated model performs over the traditional classification approach on the forum data usages. The outcomes show that student participation in the course forum is a good predictor of the result of the course and the proposed integrated approach. The highest results are obtained by naïve bays with six attributes is 89.4%.

Antonia (2008) addressed the incorporation of clustering as a complementary step to text classification and the featured representative of the texts boost the performance of SVM/TSVM classifier experiments that were carried out on ECML/PKDD discovery challenges 2008 for SPAM detection in Social bookmarking system.

Sumana (2014) They have proposed a hybrid model by cascading clustering and classification .here they have used K-means with a 10 fold cross-validation preprocessing algorithm with 12 classifiers on 5 different medical datasets. They have used a best-first search (BSF) and correlation-based feature selection (CFS) for relevant feature selection. The performance of the algorithm was measured in terms of accuracy, Kappa, mean absolute error, and Time. Experimental results showed that the proposed model with CFS and BFS as feature selection methods and with the combination of preprocessing gives the enhanced classification accuracy over medical data sets.

Yong (2009) Paper presented an improved clustering center-based classification algorithm that does not use whole training sets. Initially to eliminate the multipeak effect of the training sample sets austerity process is deployed. Training sample sets are clustered by k means clustering algorithm. Cluster centers are considered as new training samples. To measure the importance of

each sample the weight value based on their contribution to the classification results is assigned and finally, the KNN classification algorithm will be used for the classification of the inputted text. The finding showed that the proposed algorithm can not only trim down the training samples and decrease the calculation complexity but also improve the accuracy of KNN text classification algorithms.

2.4 Imbalanced Data

Kotsiantis et. al. (2006) in their review paper presented theoretical aspects of handling imbalanced data sets. They explained both data level and algorithmic level approaches. At data level they proposed random oversampling with replacements, random oversampling without replacements, directed oversampling, directed under-sampling, oversampling with the informed generation of new samples, a combination of under and oversample and Feature Selection for imbalance datasets At algorithmic level they describe cost adjusting function, adjusting decision threshold, reorganization based learning, a mixture of expert approaches.

Yen et al. (2009) proposed cluster-based under-sampling .the basic assumption behind this method is different clusters will represent different character tics .in their approach first they build several clusters out of the datasets and then select suitable number if samples from the majority class.

Yang and Wu (2007) mentioned in his article that imbalance data is one of the top 10 challenging problems in data mining.

Sumana and Santhanam (2016) presented solutions for the wider aspect of Imbalance data which comes are the hidden in every real-world datasets problem. They deal with class imbalance problems on Heart, Appendicitis, Parkinson's datasets by Smote, and ROSE methods. They eliminated class overlap by the K-means method. They deployed PCA for feature extraction. They deployed the

discarding method to deal with class overlap problems among Merging, Separating, and Discarding.

Chawla (2005) presented a novel oversampling technique, SMOTE (Synthetic Minority Oversampling Technique). He has also experiment SMOTE with Tomek and SMOTE with ENN and found that SMOTE with ENN works well with imbalanced datasets.

This paper provides a detailed review of rare event identification from imbalance learning. The study considered Five hundred and seventeen papers published in the past decades (Haixianga, 2017).

Wang et al. (2017) proposed an ensemble method named Bagging of Extrapolation Borderline-SMOTE SVM deal with imbalance data.
Rout (2018) presented a survey on all available possible solutions to deal with imbalance data they presented an extensive literature survey including methods findings and limitations.

Drown et al. (2009) presented a novel evolutionary method based on genetic algorithm for under-sampling .C 4.5 Decision Tree and RIPPER classification algorithms were used. The performance was measured using the Area under Curve-measure. It was observed that their method outperformed other sampling techniques.

Xie (2019) suggested an oversampling algorithm for the classification of imbalanced data based on the samples' selection strategy. To make the data set balanced Random-SMOTE algorithm is deployed and the Support Vector Machine is used to classify imbalanced datasets. The imbalance data sets are classified with the SVM classification algorithm. The performance of the model is measured on ten imbalanced datasets using F-measure, G-mean, ROC, and AUC.

Rahman (2013) balanced cardiovascular data using over-sampling (SMOTE) and an improved under-sampling technique. The proposed model is particularly useful for datasets where the class labels are not known.

2.5 Feature Selection

There are a few significant works done for feature selection and Extraction Thiago et al. (2011) proposed a filter-based algorithm to partition the set of features that considers feature-class correlations for feature selection. They have used the algorithm on ten datasets to illustrate the performance of the proposed algorithm; they have also presented a theoretical framework to find clusters of features.

Liu (2015) proposed a Feature Clustering with selection strategies (FECS) for Noisy software archives. They choose Eclipse and NASA software projects real-world data sets. This method has two phases the first phase is to cluster feature and the second is for feature selection in this phase three different kind of heuristic strategies are used. Strategy-1 (FR): the most relevant features are selected, Strategy- 2 (MR): The features which have maximum representatives are selected. Strategy-3 (LE): the least eccentricity is selected.

Chandrashekar (2014) gives an insight into the existing feature selection techniques falls in Filter methods or Wrapper methods or embedded methods. They have used seven datasets to demonstrate the performance on SVM and RBF on classifier accuracy and the number of reduced features to compare the feature selection techniques.

Lee (2006) proposed a new feature selection method based on information gain and divergence-based for text categorization.

Laopracha et al. (2019) proposed a novel method that uses the ideal vectors of the vehicle and non-vehicles images for selecting features in dominant patterns from the histograms of oriented gradients for the detection in vehicles. Histogram of Oriented Gradients contains ambiguous and redundant features which lead to an expensive classification for SVM, KNN, Random forest, deep neural networks with high misclassification rates. Ladha et al. (2011) have offered a feature selection that reduces dimensionality by removing irrelevant, redundant, and noisy features result in fewer storage requirements and an improved learning algorithm speed, accuracy of the model.

Fahy and Yang (2019) proposed a dynamic feature masking to cluster high dimensional data-streams. Redundant features are masked and clustering is performed on relevant unmasked features. The proposed algorithm can be easily attached to any of the density-based clustering algorithms for text stream and image stream.

Feature selection & Extraction can be applied to supervised as well as unsupervised learning (Saeys, 2007). In the proposed model selected features are applied to supervised learning (classification).

Little et al. (2008) proposed a new measure named Dysphonia, PPE (pitch period entropy) and used Support Vector Machine (Gaussian radial basis kernel functions) to classify PD dataset and reported 91.4% of accuracies and Das discussed a Neural Network classification scheme and reported 92.9% as percentage accuracies.

Luukka (2011) introduced a fuzzy entropy-based feature selection method to predict PD. The reported classification accuracy was 85.03% with only two features from the original set of features.

Li, Liu, and Hu (2011) proposed a combination of a fuzzy-based non-linear transformation method with the principal component analysis (PCA) in order to extract the optimal set of features for the SVM classifier. The best-reported classification accuracy of 93.47% was achieved.

Butterworth (2005) make use of dendrograms to select relevant features from the resulting cluster hierarchy. The technique is capable to understand the structure of the data under consideration.

Sheikhpour et al. (2017) developed feature selection methods to select relevant features from labeled and unlabeled data .they proposed two taxonomies based on the hierarchical structure of semi-supervised feature selection methods.

2.6 Research gap

There are several machine learning algorithms available to perform a specific task. A specific algorithm works on a specific type of data and fails to retain its performance with different set of data. Sometimes it becomes difficult to choose this combination. Furthermore, there are many issues related to data itself which restricts good algorithms to perform well. It is also observed in literature survey that there are several algorithms proposed by domain experts to deal with the aforementioned issues associated with data. For example, if the data is imbalanced one needs to learn a different set of algorithms that deals with imbalanced data same for high dimension, noisy or data having more missing values one has to learn several algorithms to identify the best one. This problem is addressed by Manuel Fernandez-Delgado et. al. (2014) in his paper “Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?". They quoted that “***A researcher may not be able to use classifiers arising from areas in which he/she is not an expert (for example, to develop parameter***

tuning), being often limited to use the methods within his/her domain of expertise".

There is not any single algorithm to solve these problems implicit in real-life data. To fill this gap, I have proposed an integrated framework of clustering and classification as a generic solution to deal with these challenges present in the data. In this approach, one has to learn only K means algorithm for clustering and SVM classifier for the classification task to deal with every kind of data.

2.7 Summary

After this extensive review, the potential of the integrated approach is identified and applied to resolve the critical problems in the presentation of an efficient and better model in the data-mining domain.