Evaluating Classification Techniques on Different Categories of Datasets

Sachin Singh Thakur[1] Shubham Prakash Srivastava[2]
[1]Indian Institute of Information Technology, Design & Manufacturing Jabalpur, India
[2]University of Petroleum & Energy Studies, Dehradun, India

Abstract— This paper aims to address most popular and interesting problem from the field of classification. Choosing a suitable classifier for a given problem which itself has been a challenging problem that typically requires experimentation with different classification techniques. In our work, we try to address this question by investigating any relation that may exist between the dataset characteristics and classifier bias. Specifically, in our study, we categorize datasets based on their inherent Characteristics like imbalanced, missing value, noise, and different sizes of the datasets and investigate the performance of two classes of classifiers - five Machine Learning and five Statistical Learning classifiers. Machine Learning techniques scored over Statistical techniques in five categories out of seven categories of the datasets. Multilayer Perceptron was found to be the most robust classifiers in the class of machine learning classifiers and Bayes Net was found to be the most robust classifiers in the class of Statistical Learning classifiers.

Key words: Statistical Technique, Machine Learning Technique, Area under Curve, Imbalanced Datasets, Missing Value Datasets, Noisy Datasets, Size Datasets

I. INTRODUCTION

Classification is a technique which is used to predict what categorical class a data instance belongs to [1]. The role of classification technique is to map a data item into one of the several defined classes on the basis of a training set of data whose category membership is already known. Choosing a suitable classifier for a given problem which itself has been a challenging problem that typically requires experimentation with different classification techniques. Classification techniques usually vary in their ability to cope up with the different types of datasets, and choosing the right classifier for a given application (hence the dataset) is remained as one of the tricky tasks in the field of classification.

STATLOG Project takes an initiative for comparing different classification algorithm on large real world problems [2]. Few extensive empirical studies have been carried out for comparing the classification algorithms [3][4]. According to No Free Lunch Theorem, no single classifier performs better on all type of datasets [5]. One of the fundamental problems in computer science is to choose the better classification technique among the myriad of techniques available. Some classification techniques are simple and intuitive and perform well in the small datasets. Some classification techniques handle imbalanced datasets more efficiently as compared to others. The performance of various classifiers depends upon the type of datasets being analyzed [6]. In the recent past, several authors have made an attempt for comparing classifiers [7][8][9].

We conjecture that the performance of a given classifier may influence by the inherent nature of the datasets. In our empirical study, we formulated seven categories of the datasets that are based on the missing value, noise, different sizes and imbalanced nature of the datasets. This high-level categorization of classification datasets into the relevant categories can help us to give a reason why a classifier is more suitable than others over the given category of the datasets. Our motive behind this empirical study is to look for the characteristics, in which a specific classification technique outperforms others on a defined categories of datasets.

The rest of the paper is organized as follows- Section II contains related work in the area of classification. In Section III, we present our experimental setup that includes information about project datasets we have used, metrics (in-dependent variables) and dependent variables, dataset with its multiple releases, performance measure used, machine leaning and statistical techniques used for empirical investigation. We present the methodology we followed to perform empirical investigation in Section IV and present the results of investigations in Section V. We discuss the implications of our results in Section VI. Section VII summarizes the threats to validity and Section VIII contains discussion.

II. RELATED WORK

It is very complicated to make sense of the multitude of empirical studies that have been yet. So often, results are in the contradictory state, with one author claiming that support vector machine is superior to decision tree and another making the opposite claim according to their perspective. There are no agreed subjective factors by which to judge algorithms, such as how easy an algorithm is to handle missing value datasets, such as how much effective algorithm over missing value datasets, are also very important when a researcher makes his choice from the many algorithms available [10]. Wu et al. makes some effort to identify the most influential learning algorithm. Dietterich et al. [11] efforts to design procedures to estimate the classifier performance. Demsar et al. [10] defines a statistical framework to extract reliable conclusion from results. Several researchers have attempted comparative study among classifier. Cooper et al. [3] has done the comparative study of dozen learning algorithms over medical dataset using accuracy and ROC metric. Lim et al.

[12] conducts an empirical comparison of decision trees and other classification algorithm using accuracy metric. Bauer et al. [13] presents an empirical analysis of ensemble methods such as bagging and boosting. Perlich et al. [7] conducts an empirical comparison between logistic regression and decision tree algorithms. Provost and Fawcett et al. [14] discusses the importance of evaluating learning algorithms on metrics other than accuracy such as ROC. Rich Caruana et al. [15] conducts a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps.

Despite the fact that several researchers had done research in fault prediction using machine learning technique

and statistical techniques individually [7][8][9][16], not many extensive studies have been done that compare machine learning and statistical techniques and best of them [17]. Twenty-two decision tree, nine statistical, and two neural network algorithms are compared on thirty-two datasets in terms of classification accuracy, training time, and number of leaves. The main motivation behind this study is to investigate the effectiveness of machine learning and statistical techniques to provide the guidance in selection of model building techniques. Statistics emphasize inference whereas machine learning emphasizes prediction. Zuleyka Daz et al. [18] compares three classical well known algorithms: one of them belong to the field of machine learning (Multi layer Perceptron) and two statistical techniques (Linear Discriminant Analysis and logistic regression) on a sample of Spanish nonlife insurance companies. The results demonstrated that machine learning techniques outperformed statistical techniques. STATLOG Project did the excellent comprehensive studies of different machine learning, neural and statistical classification algorithms on data sets from real- world industrial areas including medicine, finance, image analysis, and engineering design [2].

III. Experimental Setup

In this section, we present our experimental setup that includes information of the dataset, independent variable (metrics), dependent variable, machine learning techniques, statistical technique and set of hypotheses we used.

A. Performance Measure

We used four performance measures: Precision, Recall, Specificity and AUC (Area under the ROC curve) to evaluate the results of our investigation. They are define as-

Precision is the percentage of classes classified as faulty that are actually faulty, it shows, how effective we are in identifying the fault location.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Sensitivity is the percentage of faulty classes that are predicted as faulty. Recall can be also called as Sensitivity.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Specificity is the percentage of non-faulty classes that are predicted as non-faulty.

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Where, TP = True Positive is one that detects the condition when the condition is present.

- TN = True Negative is one that does not detects the condition when the condition is absent.
- FP = False Positive is one that detects the condition when the condition is absent.
- FN = False Negative is one that does not detect the condition when the condition is present.

Receiver Operating Characteristic (ROC) analysis: The performance of the output of the predicted models was evaluated using ROC analysis. It is an effective method of evaluating the performance of the model predicted. The ROC curve is defined as a plot of sensitivity on the y-coordinate versus its 1-specificity on the x-coordinate. While constructing ROC curves, we selected many cut off points between 0 and 1, and calculated sensitivity and specificity at each cut off point.

B. Classification Techniques

Classification creates a model based on which a new instance can be classified into the existing classes. Classification techniques predict an inevitable outcome based on a given input. The election of classifier technique is based on the top ten data mining algorithms [19]. Classification algorithm can be classified into two broad categories: Statistical and Machine learning techniques [18][20].

1) Statistical Techniques

Statistical Techniques is based on a probability model, which provides a probability that an instance belongs to each class, rather than classify it to one of the classes. For full details of statistical theory researcher can consult a statistical textbook written by Michie [20].

2) Logistic Regression

Logistic regression is used to model the relationship between a dependent variable and one or more independent variable. It is an approach to learning functions of the form $f : X \longrightarrow Y$, or $P(Y \mid X)$ in the case where Y is discrete-valued, and $X = X1,\{X2, \ldots, Xn \quad \}$ is any vector containing discrete or continuous variables [16].

3) Naive Bayes

Naive Bayes come in the category of statistical classifiers. They will help us to predict the class membership probabilities. The Naive Bayes work on a very intuitive concept of Bayes rule of conditional probability. Rather than prediction, Naive Bayes classifier produces probability estimates [9].

4) K-Nearest Neighbor

K-Nearest neighbor is a statistical technique works on the principle that the instances within a dataset will exist in close vicinity to other cases that have similar characteristics. K-nearest neighbor classified instance based on majority vote of its neighbor using some similarity measure like euclidean distance. The choice of k influence the classification performance [21].

5) Bayesian Networks

Bayesian networks is a probabilistic graphical model, which is used to representing knowledge about an variable domain. Specifically, each node in the graph depicts knowledge about an variable domain, while the edges between the nodes depict probabilistic dependencies among the corresponding random variables [22].

6) Multiclass LDA

Multiclass LDA is an extension of two-class LDA that can able to handle the arbitrary number of classes. Natural extension of Fisher Linear discriminant used multiple discriminant analysis, when the number of classes is more than two. In two-class, the projection is from high dimensional space to a low dimensional space and the transformation suggested still maximizes the ratio of intra-class scatter to the inter-class scatter [23].

7) Machine Learning Techniques

Machine learning technique is a learning technique which deals with making intelligent decisions based on features of the input data rather than preset rules [18][19]. In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data.

8) Support Vector Machine

Support vector machine is a machine learning technique that is based on the principle of decision planes. These decision plane is used to discriminate between a set of objects having distinct classes. It works on the principle of constructing a hyperplane or set of hyperplanes in a high or infinite dimensional plane, which can be used for classification [24].

9) Multilayer Perceptron

Multilayer perceptron is an advancement of simple perceptron that was developed to surpass the limitations that was found in the simple version of this. They are also known as feed forward networks. Multilayer Perceptron is used to solve different problems in various fields like: pattern recognition, interpolation etc. [25].

10) C4.5

The C4.5 Decision tree classifier uses a simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the training data. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain [19][26].

11) Adaboost

AdaBoost, referred as Adaptive Boosting is a machine learning meta-algorithm was introduced by Scaphire and Freund in 1997, its popular due to the modification of combining classifiers in a cascade as proposed by Viola and Jones in 2004. Adaboost boosts the classification performance by combining several weak classification functions into a stronger classifier [9].

12) CART Decision Tree

Classification and regression trees is machine-learning technique which is used to build prediction models from the data. The models are built by continuously partitioning the data space and fitting a simple prediction model within each partition [19].

C. Research Questions

In this paper, we aim to investigate the effectiveness of machine learning and statistical techniques. As a follow up, we frame our research question.

1) Research Question

Which of the machine learning and statistical techniques are more suitable for a given category of the datasets?

IV. THE EMPIRICAL STUDY

Depending on the classification technique used and datasets being analyzed, it's feasible to detect which inherent characteristics of datasets affect the performance of classification technique [27]. Since the dawn of classification, the classification problem has attracted a great deal of research [1]. Ruchika et al. has compared machine learning technique with the classical statistical algorithms [8]. Zuleyka Diaz et al.

[4] compared three classical well known algorithms: one of them belong to the field of machine learning (Multilayer Perceptron) and two statistical techniques (Linear Discriminant Analysis and logistic regression) on a sample of Spanish nonlife insurance companies. The results demonstrated that machine learning techniques outperformed statistical techniques. Empirical stud- ies have been carried out for comparing the classification algorithms [3][4]. STATLOG Project takes an initiative for comparing different classification algorithm [2]. In our pro- posed work, we evaluate the performance of two classes of classifier- five machine learning and five statistical techniques over seven defined categories of datasets.

A. Study with Imbalanced Datasets (Ctt1)

A dataset is called as imbalanced if at least one class is represented using small number of instances (minority class), while other classes make up the majority [30]. Learning from imbalanced datasets is an emerging research interest nowadays due to its significant impact on the society. A lot of efforts have been spent on resolving critical issues like detection of computer network intrusions, detection of fraudulent transactions in the bank, and detection of cancer contributing genes that have arisen due to the imbalanced dataset [28].

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to thirty standard imbalanced datasets collected from the UCI collection. These datasets contained the number of attributes, number of instances, number of classes and the imbalanced ratio. Imbalanced ratio lies from 1.82 to 10. We have used a wide range of datasets in our experiments as summarized in Table 3.2. The number of features varies from 4 to 20; the number of instances ranges from 150 to 5472 and the number of classes is 2.

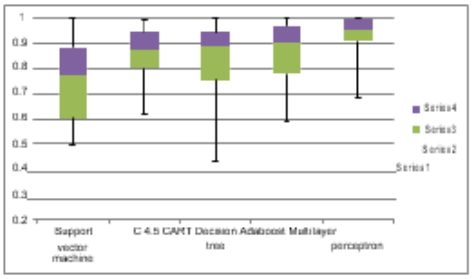

Fig. 1: Imbalanced Datasets (a) Box Plot of AUC for Machine Learning Techniques

2) Results & Analysis

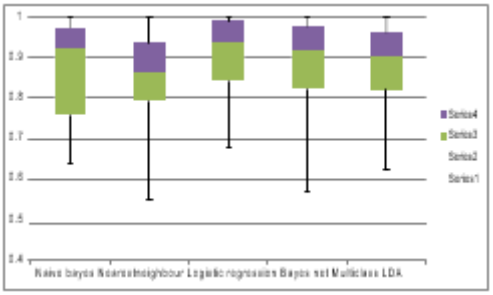

Fig. 2: Imbalanced Datasets (b) Box Plot of AUC for Statistical Techniques

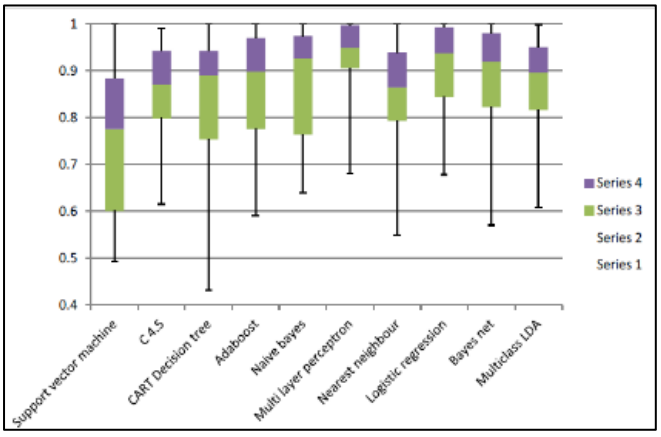

Fig. 3: Imbalanced Datasets (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on twenty-eight imbalanced datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in imbalanced datasets. For the main comparison, Figure 3.2 contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

- RQ 1: We conclude from box plot diagram that that Multilayer perceptron performs better among the machine learning techniques.
- RQ 2: Logistic Regression performs better among statistical techniques.
- RQ 3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that machine learning techniques score over statistical techniques.

B. Study with Missing value Datasets (Ctt2)

Many research datasets contain missing value due to the incorrect measurements, manual data entry error, and fault due to the equipment readings [30]. Dataset which incorporates missing value adversely affect the performance of a classifier. Issues associated with missing values are a loss of efficiency and bias resulting from difference between missing value and complete data. The presence of missing values in a dataset can affect the performance of the classifier. The missing value is a common problem in statistical analysis. Rates of less than 1% missing value in dataset are trivial. If the range lies between 1- 5%, it is manageable. However, when range increases from 5%, some sophisticated method must be required to handle missing value dataset. Majority of research in the data mining community based on the performance of classification algorithms on how well they perform on missing value datasets [31][32]. Arff files used "?" symbol for missing values. Missing value in a dataset can affect the performance of classifier which leads to the difficulty of extracting relevant information from the dataset. To estimate the performance of classifier we have used AUC performance measure. WEKA tool is used for this study.

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to thirty standard missing value datasets from the UCI collection. These datasets contained the number of attributes, number of instances, number of classes and missing value ratio. Missing value ratio lies within a range of 2.29% to 48.39%. We have used a wide range of datasets in our experiments as summarized in Table.

The number of features ranges from 5 to 41, the number of instances ranges from 33 to 299284, and the number of classes varies from 2 to 22. We consider the diversity of datasets that contains a variable number of features, number of instances, and missing value ratio.

2) Results & Analysis

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on thirty missing value datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in imbalanced datasets. For the main comparison, Figure 3.3 contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

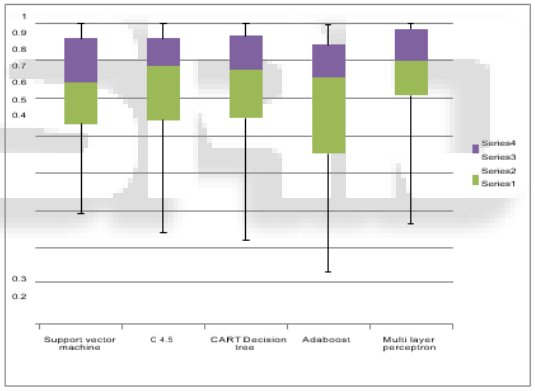

Fig. 4: Missing value Datasets (a) Box Plot of AUC for Machine Learning Techniques

- RQ 3.1: We conclude from the box plot diagram that that Multilayer Perceptron performs better among the machine

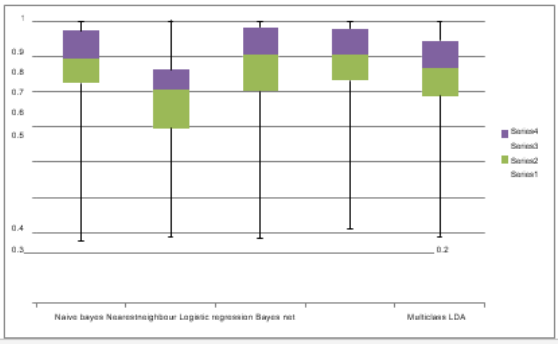

Fig. 5: Missing value Datasets (b) Box Plot of AUC for Statistical Techniques learning techniques and Support Vector Machine performs worst among the machine learning techniques.

– RQ 3.2: Bayes Net and Naive Bayes performs better among the statistical techniques while nearest neighbor performs worst among the statistical techniques.
– RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that machine learning techniques score over statistical techniques.

C. Study with Small Number of Features & Small Number of Instances (Ctt3) Datasets

Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis. Dataset that has number of features (p>=20) and number of cases (n>= 1000) is categorized as a large dataset. Dataset that has number of features (p<=7) and number of cases (n<=500) is categorized as the small dataset. These two categories are combined with each other, and formulated four categories of datasets and then conduct a study to learn the behavior of the machine learning and statistical techniques using it. This category of datasets contains number of features ($p \leq 7$) and number of cases ($n \leq 500$).

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to the thirty *Ctt*3 datasets. These datasets contained the number of attributes, number of instances, and number of classes. We have used a wide range of datasets in our experiments as summarized in Table 3.4. The number of features ranges from 3 to 7, the number of instances ranges from 15 to 490 and the number of classes ranges from 2 to 9. Challagulla et al. [33] uses some of the characteristics of the software data sets that we are using in our analysis. If the number of cases (n 500) and number of features (p 7) then that dataset is considered as Small number of features and small number of instances (*Ctt*3).

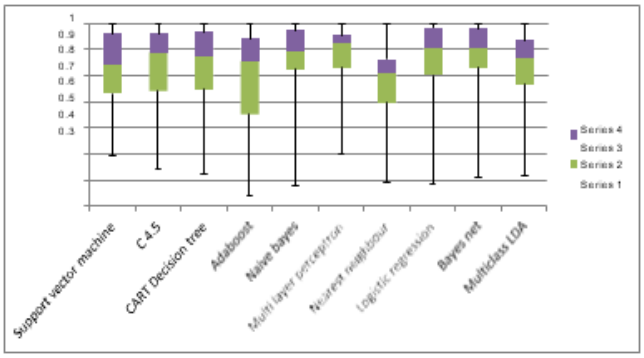

Fig. 6: Missing value Datasets (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques learning techniques and Support Vector Machine performs worst among the machine learning techniques.

– RQ 3.2: Bayes Net and Naive Bayes performs better among the statistical techniques while Nearest neighbor performs worst among the statistical techniques.
– RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that machine learning techniques score over statistical techniques.

D. Study with Small Number of Features & Small Number of Instances (Ctt3) Datasets

Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis. Dataset that has number of features (p>=20) and number of cases (n>= 1000) is categorized as a large dataset. Dataset that has number of features (p<=7) and number of cases (n<=500) is categorized as the small dataset. These two categories are combined with each other, and formulated four categories of datasets and then conduct a study to learn the behavior of the machine learning and statistical techniques using it. This category of datasets contains number of features ($p \leq 7$) and number of cases ($n \leq 500$).

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to the thirty *Ctt*3 datasets. These datasets contained the number of attributes, number of instances, and number of classes. We have used a wide range of datasets in our experiments as summarized in Table 3.4. The number of features ranges from 3 to 7, the number of instances ranges from 15 to 490 and the number of classes ranges from 2 to 9. Challagulla et al. [33] uses some of the characteristics of the software data sets that we are using in our analysis. If the number of cases (n 500) and number of features (p 7) then that dataset is considered as Small number of features and small number of instances (*Ctt*3).

2) Results & Analysis

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on thirty (*Ctt*3) datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in *Ctt*3 datasets. For the main comparison, Figure contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

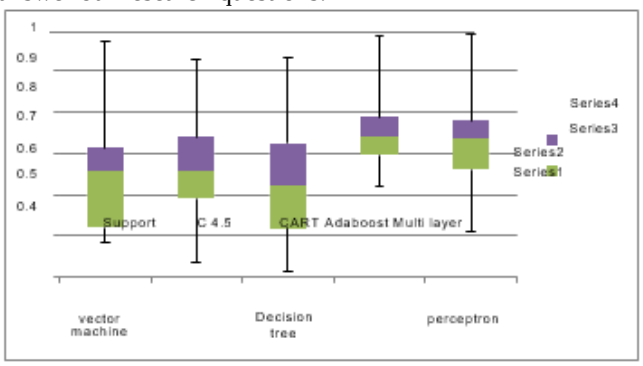

Fig. 7: Small Number of Features and Small Number of Instances (a) Box Plot of AUC for Machine Learning Techniques

From the box plot comparison:

– RQ 3.1: We conclude from the box plot diagram that that Adaboost performs better among the machine learning techniques and CART Decision Tree performs worst among the machine learning techniques.

- RQ 3.2: Logistic Regression performs better among the statistical techniques while Nearest Neighbor performs worst among the statistical techniques.
- RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that machine learning techniques score over statistical techniques.

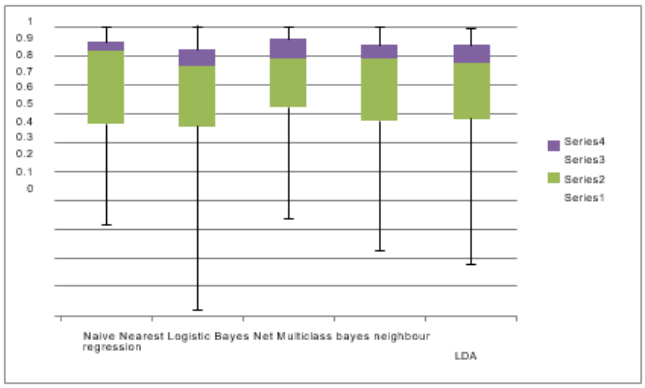

Fig. 8. Small Number of Features and Small Number of Instances (b) Box Plot of AUC for Statistical Techniques

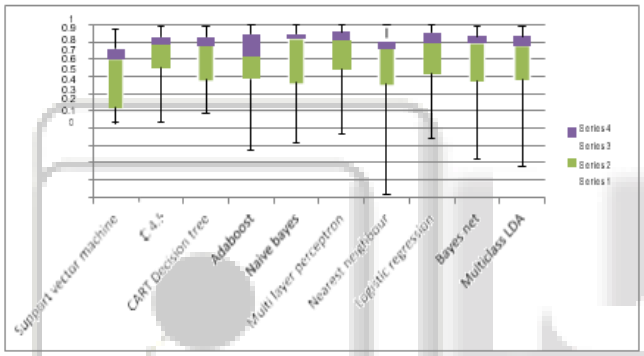

Fig. 9: Small Number of Features & Small Number of Instances (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques

E. Study with Small Number of Features & Large Number of Instances (Ctt4) Datasets

Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis. Datasets which has number of features (p>=20) and number of cases (n >=1000) is considered as the large dataset. Dataset which has number of features (p<=7) and number of cases (n<=500) is considered as the small dataset. These two categories are combined with each other, and formulated four categories of datasets and then conduct a study to learn the behavior of the machine learning and statistical techniques using it. This category of datasets contain number of features ($p \leq 7$) and number of cases ($n \geq 1000$).

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to the thirty small number of features and large number of instances (*Ctt*4) datasets from the UCI collection. These datasets and their characteristics are summarized in Table 3.3. These datasets contained the number of attributes, number of instances, and number of classes. We have used a wide range of datasets in our experiments as summarized in Table 3.5. The number of features ranges from 20 to 93, the number of instances ranges from 27 to 440. Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis: number of cases (n 500) and number of features (p 20). Dataset which has these characteristics is considered as Large number of features and small number of instances (*Ctt*4).

2) Results and Analysis

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on thirty (*Ctt*4) datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in *Ctt*4 datasets. For the main comparison, Figure contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

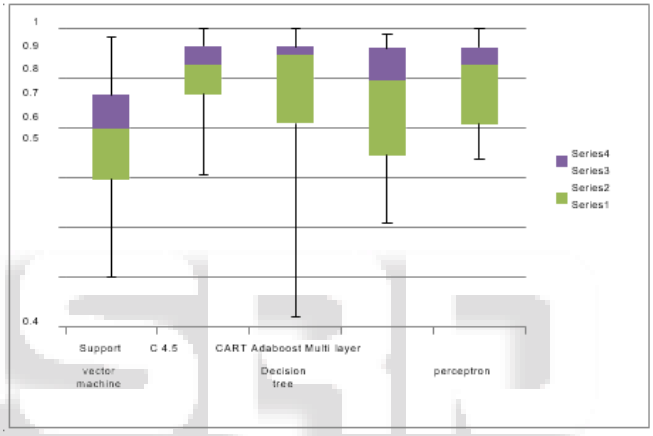

Fig. 10: Small Number of Features & Large Number of Instances (A) Box Plot of AUC for Machine Learning Techniques

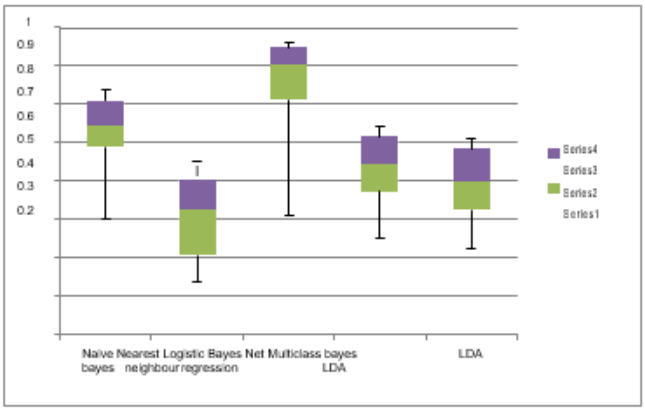

Fig. 11: Small Number of Features & Large Number of Instances (B) Box Plot of AUC for Statistical Techniques

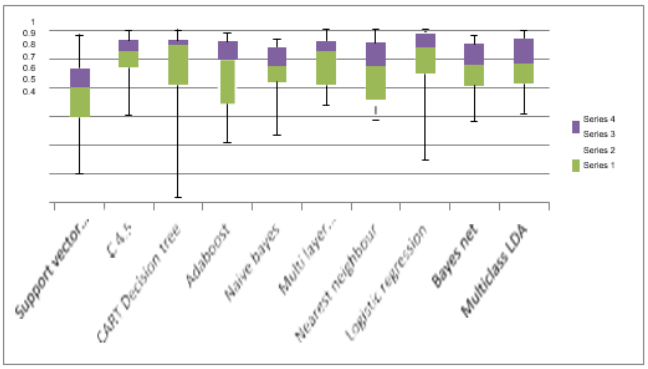

Fig. 12: Small Number of Features & Large Number of Instances (C) Box Plot of AUC for Comparing Machine Learning & Statistical Techniques

- RQ 3.1: We conclude from the box plot diagram that that CART Decision Tree performs better among the machine learning techniques and C4.5 performs worst among the machine learning techniques.
- RQ 3.2: Logistic Regression performs better among the statistical techniques while Nearest Neighbor performs worst among the statistical techniques.
- RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that statistical techniques score over machine learning techniques.

F. Study with Large number of Features & Small number of Instances (Ctt5) Datasets

Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis: Dataset which has number of features (p>=20) and number of cases (n>=1000) is considered as the large dataset. Dataset which has number of features (p<=7) and number of cases (n<=500) is considered as the small dataset. These two categories are combined with each other, and formulated four categories of datasets and then conduct a study to learn the behavior of the machine learning and statistical techniques using it. This category of datasets contain number of features ($p \geq 20$) and number of cases ($n \geq 1000$).

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [28][29]. Our experiment applied the five machine learning and five statistical techniques to thirty standard datasets from the UCI collection. These datasets and their characteristics are summarized in Table 3.6. These datasets contained number of attributes, number of instances and number of classes. We have used a wide range of datasets in our experiments as summarized in table I. The number of features ranges from 20 to 93, the number of instances ranges from 27 to 440. Challagulla et al. [33] uses some of the characteristics of the software data sets that we are using in our analysis: number of cases ($n \leq 500$) and number of features (p ≥20) Dataset which have these characteristics is considered as Large number of features and small number of instances (*Ctt*5).

2) Results & Analysis

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on thirty (*Ctt*5) datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in *Ctt*5 datasets. For the main comparison, Figure contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

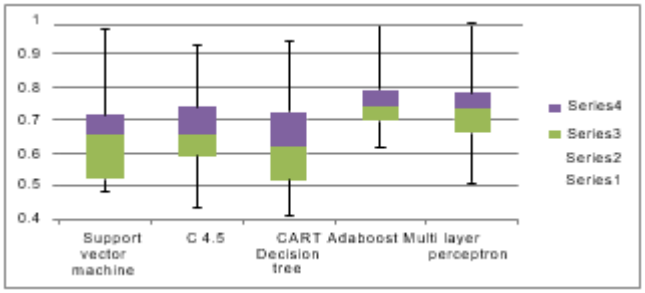

Fig. 13: Large number of Features and Small number of Instances (a) Box Plot of AUC for Machine Learning Techniques

- RQ 3.1: We conclude from the box plot diagram that that Adaboost performs better among the machine learning techniques and CART Decision Tree performs worst among the machine learning techniques.
- RQ 3.2: Naive Bayes performs better among the statistical techniques while Nearest Neighbor performs worst among the statistical techniques.
- RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box

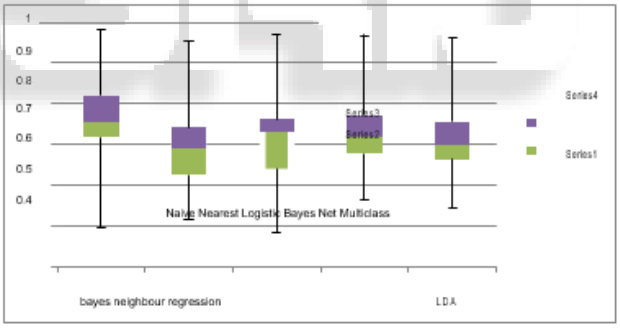

Fig. 14: Large number of Features and Small number of Instances (b) Box Plot of AUC for Statistical Techniques

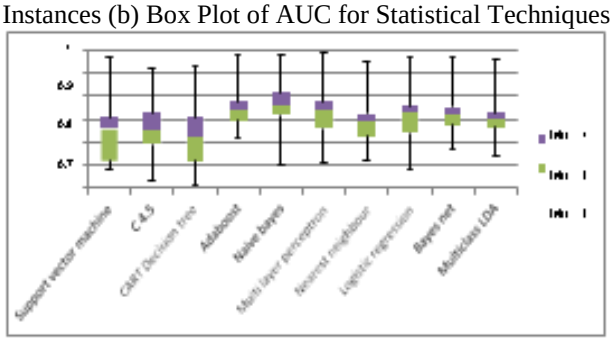

Fig. 15: Large number of Features and Small number of Instances (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques

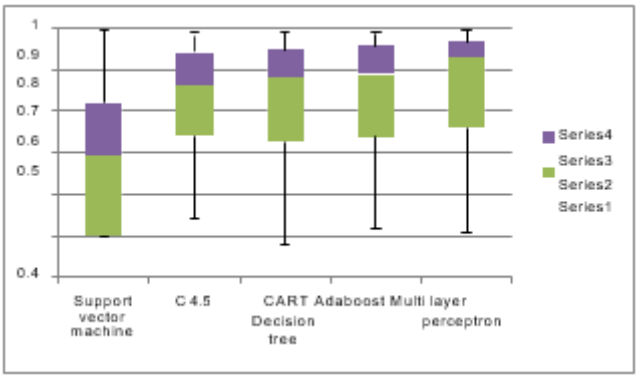

Fig. 16: Large Number of Features & Large number of Instances (a) Box Plot of AUC for Machine Learning Techniques plot diagram that machine learning techniques score over statistical techniques.

G. Study with Large number of Features & Large number of Instances (Ctt6) Datasets

Challagulla et al. [33] use some of the characteristics of the software data sets that we are using in our analysis: Dataset which has number of features (p>=20) and number of cases (n >=1000) is considered as the large dataset. Dataset which has number of features (p<=7) and number of cases (n <=500) is considered as the small dataset. These two categories are combined with each other, and formulated four categories of datasets and then conduct a study to learn the behavior of the machine learning and statistical techniques using it. This category of datasets contain number of features ($p \geq 20$) and number of cases ($n \geq 1000$).

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine learning and five statistical techniques to thirty standard *Ctt*6 datasets. These datasets and their characteristics are summarized in Table 3.7. These datasets contained number of attributes, number of instances, and number of classes. The number of features ranges from 20 to 217, the number of instances ranges from 1050 to 23014. Challagulla et. al. [33] uses some of the characteristics of the software data sets that we are using in our analysis: number of cases (n >=1000) and number of features (p>=20). Dataset which have these characteristics is considered as Large number of features and small number of instances (*Ctt*6).

2) Results & Analysis

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on thirty (*Ctt*6) datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in *Ctt*6 datasets. For the main comparison, Figure contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

- RQ 3.1: We conclude from the box plot diagram that that Multilayer Perceptron performs better among the machine learning techniques and Support Vector Machine performs worst among the machine learning techniques.
- RQ 3.2: Logistic Regression performs better among the statistical techniques while Naive Bayes performs worst among the statistical techniques.
- RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that statistical techniques score over machine learning techniques.

H. Study with Noisy Datasets (Ctt7)

Learning classifiers in the presence of label noise is a classical problem in machine learning [34]. It is the most challenging problem in machine learning. This issue has been approached from different direction by different researchers. A detailed survey of these approaches is discussed by Frenay

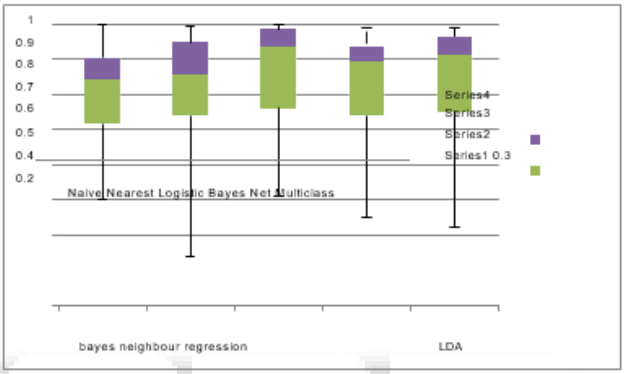

Fig. 17: Large Number of Features and Large number of Instances (b) Box Plot of AUC for Statistical Techniques

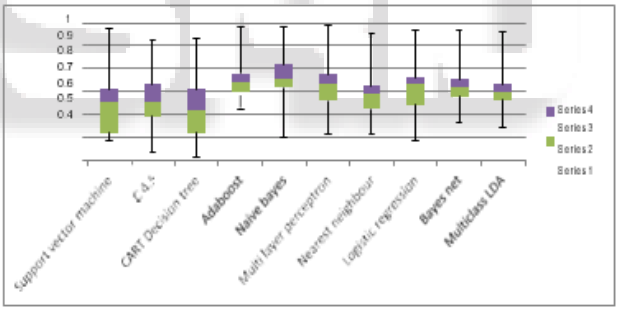

Fig. 18: Large Number of Features and Large number of Instances (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques

Verleysen [36]. Nettleton et al. [35] present an empirical investigation of robustness of many standard classifier learning techniques to noise in training data. They come with a finding that naive bayes has the better noise tolerance properties. There are various strategies for dealing with noisy datasets. In general, the methods for treating noisy data can be partitioned into two categories in the first approach data is preprocessed to clean the noisy points and then the standard algorithm is applied to it. In the second approach learning algorithm itself is designed in such a manner that noise doesn't affect the algorithm. When noise is already present, it can be removed by using one of the following methods: manual inspection with the use of predefined constraints on feature values, binning and clustering [37].

1) Project Datasets

The datasets we have used in this study is collected from UCI Repository [29]. Our experiment applied the five machine

learning and five statistical techniques to standard nineteen noisy datasets. These datasets and their characteristics are summarized in Table 3.7. Noisy datasets available with 5%, 10%, 15%, 20% of noise in training and test sets. These datasets and their characteristics are summarized in Table 3.5. These datasets contained number of attributes, number of instances and the number of classes. The number of features ranges from 5 to 41, the number of instances ranges from 33 to 299284 and the number of classes ranges from 2 to 22. We consider the considerable diversity of datasets in characteristics that are the number of features and the number of instances.

2) Results & Analysis

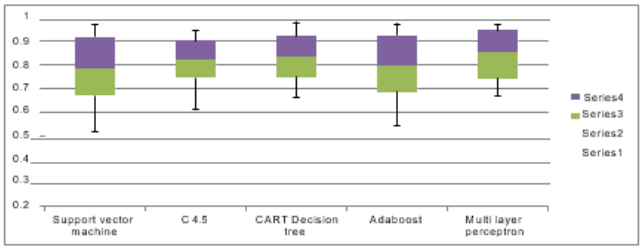

Fig. 19: Noisy Datasets (a) Box Plot of AUC for Machine Learning Techniques

In this section, we investigated the results obtained by applying five machine learning and five statistical techniques on nineteen noisy datasets. These experiments help us to state which machine learning and statistical technique among the five machine learning and five statistical techniques lead to better result in noisy datasets. For the main comparison, Figure 3.8 contains three panels: Box plot of AUC for machine learning techniques, Box plot of AUC for statistical techniques and Box plot of AUC for comparing machine learning and statistical techniques. From the Boxplot comparison, we answer our research questions:

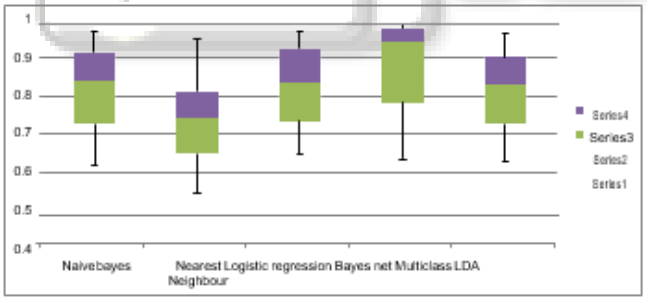

Fig. 20: Noisy Datasets (b) Box Plot of AUC for Statistical Techniques

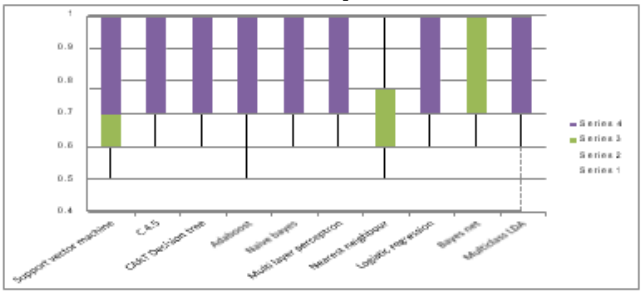

Fig. 21: Noisy Datasets (c) Box Plot of AUC for Comparing Machine Learning and Statistical Techniques

– RQ 3.1: We conclude from the box plot diagram that that Multilayer Perceptron seems to be more robust as the attribute noise increased among five machine learning techniques while CART Decision Tree performs worst as the attribute noise increased among machine learning techniques.
– RQ 3.2: Bayes Net performs better among the statistical techniques while Nearest Neighbor performs worst among the statistical techniques.
– RQ 3.3: We compare the best machine learning technique with the best statistical technique and conclude from the box plot diagram, that statistical techniques score over machine learning techniques.

V. DISCUSSION

Classification techniques vary in their ability to cope up with the different types of dataset. Choosing a suitable classifier for a given application is one of the tricky task. There is a specific technique that significantly outperformed others in particular category of datasets; this study gives the direction of choosing the same. Several techniques presented in our empirical study claims to be able to improve classification performance over certain benchmarks of datasets. This study raises an important question: why a classifier is more suitable than the others over the given category of the datasets.

A. Missing Value Datasets

We investigated the missing value tolerant capability of ten classification techniques (five machine learning and five statistical techniques). Issues associated with missing values are a loss of efficiency and bias resulting from different between missing value and complete data. Our results demonstrated that Multilayer Perceptron performs better among the machine learning techniques. There are special facilities in multilayer perceptron to handle missing values. Numeric data is scaled into the appropriate range for the network, and missing values can be substituted for using the mean value of that variable across the other available training cases.

On the other hand, Naive Bayes performs better among the statistical techniques used in our empirical study of missing value datasets. Naive Bayes is a robust to missing values, since these are simply neglected in computing probabilities and hence no impact on the final decision of classification.

B. Noisy Datasets

We investigated the noise tolerant capability of ten classification techniques (five machine learning and five statistical techniques). Learning classifiers in the presence of label noise is a classical problem in machine learning. We investigate the noise tolerant capability of 10 classification techniques (five machine learning and 5 statistical techniques). These techniques don't use any preprocessing of the data, but the technique is designed in such a way that its output is not affected much by the label noise in the training data. Our results demonstrated that Multilayer Perceptron performs better among the machine learning techniques. Multilayer Perceptron contains various hidden layers that cause the knowledge representation to be distributed throughout the network in a way that increases noise tolerance. Noisy points can frequently participate in updating the hyperplane in updating the hyperplane parameters in the Multi perceptron algorithm, as noisy points are tough to be

correctly classified, Thus, allowing a negative margin around the classification boundary can avoid frequent hyperplane updates caused due to the misclassifications with small margin. Putting an upper bound on the number of mistakes allowed for any example also controls the effect of label noise.

On the other hand, Bayes Net performs better among the statistical techniques used in our empirical study of noisy datasets. Bayes Net is a noise tolerant algorithm in our study. Bayes learner which uses conditional probabilities to derive posterior probabilities. As conditional probability values are relatively less sensitive to data errors, this would lead us to expect that Bayes Net would perform more favorably with the other chosen learning methods.

C. Imbalanced Datasets

We investigated the nature of ten classification techniques (five machine learning and five statistical techniques) over imbalanced datasets. In the conventional learning strategy, the majority class becomes a huge distraction and impedes the learning of the minority class, when learning from the imbalanced datasets. Our results demonstrated that Multilayer Perceptron performs better among the machine learning techniques. Multilayer perceptron is not sensitive to the class imbalance problem when applied to linearly separable domains, its sensitivity increases with the complexity of the domain. The size of the training set does not appear to be a factor.

On the other hand, Logistic Regression performs better among the statistical techniques used in our empirical study of imbalanced datasets. In model construction using logistic regression, imbalanced dataset is not at an issue. For logistic regression specifically, there was absolutely no benefit to creating a balanced sample. What is more important is using all the data that are available. For example, for a voting campaign, if you had 500 responses and 25,000 non-responses you got better models by using all 25,500 cases, compared to sampling down the non-responses to 550 or by weighting up the 550 responses.

D. Robust Classifiers

In our empirical study of ten classification techniques, the most robust classification techniques are Multilayer Perceptron and Bayes Net. Multilayer Perceptron is the robust classification technique among the five machine learning techniques. Multilayer Perceptron is capable of handling noise, missing value and imbalanced nature of the dataset. Multilayer Perceptron contains various hidden layers that cause the knowledge representation to be distributed throughout the network in a way that increases noise tolerance. In Multilayer Perceptron, numeric data is scaled into the appropriate range for the network, and missing values can be substituted for using the mean value of that variable across the other available training cases. Multilayer Perceptron is not sensitive to the class imbalance problem when applied to linearly separable domains, its sensitivity increases with the complexity of the domain. The size of the training set does not appear to be a factor.

Bayes Net is the robust classification technique among the five statistical techniques. Bayes Net is a noise tolerant technique. Bayes learner which uses conditional probabilities to derive posterior probabilities. As conditional probability values are relatively less sensitive to data errors, this would lead us to expect that Bayes Net would perform more favorably with the other chosen learning methods. In the Bayesian network framework, missing data is marginalized out by integrating over all the possibilities of the missing values.

VI. THREATS TO VALIDITY

There are a number of potential threats to the validity of our study. For instance, we investigated the comparative performance of machine learning and statistical techniques on limited datasets. Our models are build and evaluated on datasets available in public data repositories. Second, being a large field of research, new classification techniques are continuously being proposed. We used a small subset of the machine learning and statistical techniques but we are confident that our subset is a representative one, being based on techniques which have different modelling mechanism and are currently active fields of research.

VII. CONCLUSION & FUTURE WORK

In this investigation, we tried to address this question by investigating any relation that may exist between the dataset characteristics and classifier bias. Specifically, in our study, we categorized datasets based on their inherent characteristics like imbalanced, missing value, noise, and different sizes of the datasets and investigate the performance of ten popular classifiers. This high-level categorization of classification datasets into the relevant categories can help us to give a reason why a classifier is more suitable than others over the given category of the datasets. Machine learning techniques scored over statistical techniques in five categories out of seven categories of the datasets. Our results suggest that Multilayer Perceptron was found to be the most robust classifiers in the class of machine learning classifiers, whereas Bayes Net was found to be the most robust classifiers in the class of Statistical Learning classifiers.

We intend to look up for some alternative approaches to investigate and validate our results. Replicating the study on additional datasets may help strengthening the findings further.

REFERENCES

[1] T. N. Phyu, Survey of classification techniques in data mining, in Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, 2009, pp. 1820.
[2] R. D. King, C. Feng, and A. Sutherland, Statlog: comparison of clas- sification algorithms on large real-world problems, Applied Artificial Intelligence an International Journal, vol. 9, no. 3, pp. 289333, 1995.
[3] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa et al., An evaluation of machine-learning methods for predicting pneumonia mortality, Artificial intelligence in medicine, vol. 9, no. 2, pp. 107138, 1997.

[4] Z. Diaz, M. J. Segovia, J. Fernandez et al., Machine learning and statistical techniques: an application to the prediction of insolvency in Spanish non-life insurance companies, 2005.
[5] D. Wolpert, No free lunch theorem for optimization, IEEE Transactions on Evolutionary Computation, no. 1, pp. 467482, 1997.
[6] L. Rendell and H. Cho, Empirical learning as a function of concept character, Machine Learning, vol. 5, no. 3, pp. 267298, 1990.
[7] C. Perlich, F. Provost, and J. S. Simon off, Tree induction vs. logistic regression: A learning-curve analysis, The Journal of Machine Learning Research, vol. 4, pp. 211255, 2003.
[8] Kaur and R. Malhotra, Application of random forest in predicting fault-prone classes, in Advanced Computer Theory and Engineering, 2008. ICACTE08. International Conference on. IEEE, 2008, pp. 3743.
[9] J. P. Vink and G. de Haan, Comparison of machine learning techniques for target detection, Artificial Intelligence Review, vol. 43, no. 1, pp. 125139, 2015.
[10] J. Demsar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research, vol. 7, pp. 130, 2006.
[11] T. G. Dietterich, Ensemble methods in machine learning, in Multiple classifier systems. Springer, 2000, pp. 115.
[12] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine learning, vol. 40, no. 3, pp. 203228, 2000.
[13] E. Bauer and R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, Machine learning, vol. 36, no.1-2, pp. 105139, 1999.
[14] F. Provost and P. Domingos, Tree induction for probability-based ranking, Machine Learning, vol. 52, no. 3, pp. 199215, 2003.
[15] R. Caruana and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 161168.
[16] N. Ohlsson, M. Zhao, and M. Helander, Application of multivariate analysis for software fault prediction, Software Quality Journal, vol. 7, no. 1, pp. 5166, 1998.
[17] R. Malhotra and A. Jain, Fault prediction using statistical and machine learning methods for improving software quality, Journal of Information Processing Systems, vol. 8, no. 2, pp. 241262, 2012.
[18] Z. Diaz, M. J. Segovia, J. Fernandez et al., Machine learning and statistical techniques: an application to the prediction of insolvency in Spanish non-life insurance companies, 2005.
[19] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip et al., Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no. 1, pp. 137, 2008.
[20] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Machine learning, neural and statistical classification, 1994.
[21] L. E. Peterson, K-nearest neighbor, Scholarpedia, vol. 4, no. 2, p. 1883, 2009.
[22] F. V. Jensen, An introduction to Bayesian networks. UCL press London, 1996, vol. 210.
[23] M. W. Knuiman, H. T. Vu, and M. R. Segal, An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease, Journal of cardiovascular risk, vol. 4, no. 2, pp. 127134, 1997.
[24] J. Burges, A tutorial on support vector machines for pattern recog- nition, Data mining and knowledge discovery, vol. 2, no. 2, pp. 121167, 1998.
[25] T. M. Khoshgoftaar, E. B. Allen, J. P. Hudepohl, and S. J. Aud, Application of neural networks to software quality modeling of a very large telecommunications system, Neural Networks, IEEE Transactions on, vol. 8, no. 4, pp. 902909, 1997.
[26] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.
[27] L. Rendell and H. Cho, Empirical learning as a function of concept character, Machine Learning, vol. 5, no. 3, pp. 267298, 1990.
[28] G. M. Weiss, Mining with rarity: a unifying framework, ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 719, 2004.
[29] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/, 2010.
[30] E. Acuna and C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in Classification, Clustering, and Data Mining Applications. Springer, 2004, pp. 639647.
[31] P. K. Sharpe and R. Solly, Dealing with missing values in neural network based diagnostic systems, Neural Computing & Applications, vol. 3, no. 2,pp. 7377, 1995.
[32] P. Juszczak and R. P. Duin, Combining one-class classifiers to classify missing data, in Multiple Classifier Systems. Springer, 2004, pp. 92101.
[33] V. U. B. Challagulla, F. B. Bastani, I.-L. Yen, and R. A. Paul, Empirical assessment of machine learning based software defect prediction techniques, International Journal on Artificial Intelligence Tools, vol. 17, no. 02, pp. 389 400, 2008.
[34] P. J. Huber, Robust statistics. Springer, 2011.
[35] F. Nettleton, A. Orriols-Puig, and A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artificial intelligence review, vol. 33, no. 4, pp. 275306, 2010.
[36] B. Frenay and M. Verleysen, Classification in the presence of label noise: a survey, Neural Networks and Learning Systems, IEEE Transactions on, vol. 25, no. 5, pp. 845869, 2014.
[37] J. Van Hulse and T. Khoshgoftaar, Knowledge discovery from imbalanced and noisy data, Data & Knowledge Engineering, vol. 68, no. 12, pp. 1513 1542, 2009.

