IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2019, accepted January 21, 2019, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899578

Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data

SOLOMON H. EBENUWA[1], MHD SAEED SHARIF[1], MAMOUN ALAZAB[2], (Senior Member, IEEE), AND AMEER AL-NEMRAT[1]

[1]School of Architecture Computing and Engineering, University of East London, London E16 2RD, U.K.
[2]College of Engineering, IT & Environment, Charles Darwin University, Casuarina, NT 0810, Australia

Corresponding author: Solomon H. Ebenuwa (u0744306@uel.ac.uk)

ABSTRACT Data are being generated and used to support all aspects of healthcare provision, from policy formation to the delivery of primary care services. Particularly, with the change of emphasis from curative to preventive medicine, the importance of data-based research such as data mining and machine learning has emphasized the issues of class distributions in datasets. In typical predictive modeling, the inability to effectively address a class imbalance in a real-life dataset is an important shortcoming of the existing machine learning algorithms. Most algorithms assume a balanced class in their design, resulting in poor performance in predicting the minority target class. Ironically, the minority target class is usually the focus in predicting processes. The misclassification of the minority target class has resulted in serious consequences in detecting chronic diseases and detecting fraud and intrusion where positive cases are erroneously predicted as not positive. This paper presents a new attribute selection technique called variance ranking for handling imbalance class problems in a dataset. The results obtained were compared to two well-known attribute selection techniques: the Pearson correlation and information gain technique. This paper uses a novel similarity measurement technique ranked order similarity-ROS to evaluate the variance ranking attribute selection compared to the Pearson correlations and information gain. Further validation was carried out using three binary classifications: logistic regression, support vector machine, and decision tree. The proposed variance ranking and ranked order similarity techniques showed better results than the benchmarks. The ROS technique provided an excellent means of grading and measuring the similarities where other similarity measurement techniques were inadequate or not applicable.

INDEX TERMS Imbalanced dataset, class distribution, binary class, imbalance ratio, majority class, minority class, oversampling, under sampling, logistic regression, support vector machine, decision tree, ranked order similarity, peak threshold accuracy.

I. INTRODUCTION

The problems associated with class imbalance are very common in real-life datasets, which could result in the sensitivities of predictions becoming skewed towards the majority class target [1], while adversely becoming insensitive to the minority target class; hence, the proportions of the captured minority target classes are subjective and, in many instances, mere approximations. In data mining and other predictive scenarios, the minority classes are usually the interest group, for instance, medical data such as diabetes or heart data [2], financial fraud detection, credit scoring [3], Weblogs, and instruction detection, the interest groups are usually the minority class. Therefore, the general performance of any classification algorithm is relatively determined by the sensitivities to the minority target class, but usually, the predictions of the minority target class are below optimal due to the primary design of the algorithms, which assume an equal class distribution in both concept and application [4]. The effects of this class imbalance become more evident when the built model is applied to test data or deployed. External effects have more influence on the imbalanced data; missing data or general noise has more impact on a data distribution that is imbalanced than those that are more closely balanced; the more a dataset is imbalanced, the greater the impact of noise on the model built using the data [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

An imbalanced dataset has some unexpected impacts directly and indirectly on the deployments due to the class distributions. It is relatively difficult to assess the accuracy of the majority class that has been captured by the predictions because the accuracy boundaries are not clearly defined [5] which raises the issues of subjective proportions. It should be clear what a data mining or machine learning process is intended to achieve. In some processes, it is costlier to misclassify a minority class than to misclassify any of the majority classes, for example, in chronic diseases such as diabetes, heart disease, kidney disease and cancer [6]. If the prediction is unable to identify the sick patient, the error could result in death or more serious health complications [7].

Class imbalance also has a profound effect on big data (extremely large datasets) such that the traditional techniques of analyzing these voluminous datasets cannot produce the expected accurate results [8]. Many of the poor results in the classification performance are due to skewness in the imbalanced data class ratio. The general model built with imbalanced data has a tendency of false alarms and, in many instances, outright misclassification.

Binary classifications (two classes, 0, 1) [9], [10] are very common in imbalanced classed scenario and provides a context example of the problems encountered during predictions. [11]–[13]. An analogy is to consider a study of a population consisting of 1000 patients, and assuming that 900 patients out of 1000 have no disease as shown in Figure 1, a model that predicts all 1000 as not having the disease would still appear to be 90% accurate, even if the remaining 100 patients have the disease, and they were not identified [14].

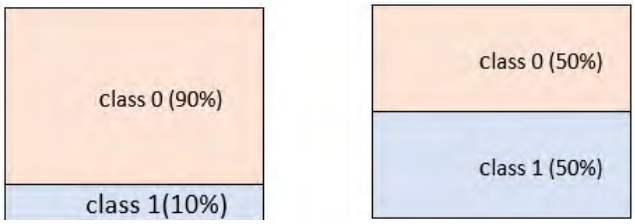

FIGURE 1. (a) Imbalance data and (b) balance data.

Therefore, the challenge is to reduce the effects of the imbalance ratio to improve the abilities of the classification algorithm to become more sensitive to the minority class because the minority classes are usually the reason for the predictions [15]. Hence, if the algorithm used is unable to perform by targeting the minority class, the accuracy obtained becomes subjective and essentially an approximation. This paper presents a new attribute selection technique (variance ranking) for handling imbalanced class problems in the dataset, which are based on the intrinsic characteristics of the data items (the variance). The contributions of this paper are as follows:

- A novel attribute selection technique (variance ranking) based on the intrinsic properties of each attribute in the subsection of the classes.
- A novel similarity measurement technique (ranked order Similarity-ROS) in Section IV-B.
- A novel method of choosing significant attributes based on ranking by showing that the significant attributes are those at the peak threshold performance.
- An independent process of evaluating and validating the proposed approach and findings.

The remainder of this paper is organized as follows: Section II provides an overview of the related works, and Section III describes the proposed method and approach. Section IV provides the experimental results based on the attribute selection on a set of publicly available data to validate the proposed framework, and Section V is the discussion, summary of the papers novelty and the conclusion.

II. RELATED WORK

Most of the real-life datasets are imbalanced, where the classes are not evenly distributed, Figure 2 represents the ubiquitous nature of imbalanced classed in association with other problems that a real life data might have; that is "Most Real-life data set must have imbalanced classed in addition to other problems". Therefore, how to obtain an accurate prediction from such a dataset is a subject of research in industry and academia. Researchers have been devising ways to reduce the general effects of class imbalance and improve the predictive performance of classification algorithms.

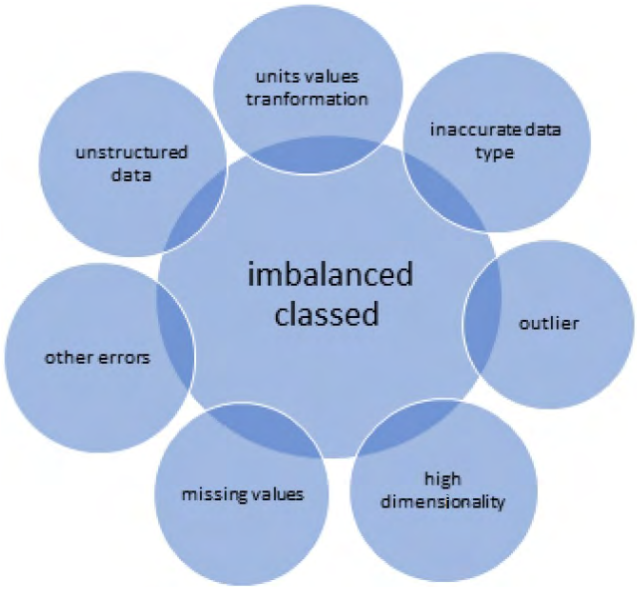

FIGURE 2. Problems of real-life data sets.

The processes of dealing with imbalanced data can be categorized as follows: *sampling-based techniques, algorithm modifications and the cost-Sensitive Approach* [16]. Details of these techniques and their importance in dealing with class imbalance and targeting the minority class will be reviewed in the following sections.

A. SAMPLING BASED TECHNIQUES

The main idea behind sampling-based techniques is to balance the class distribution. This method of handling imbalanced data has become one of the most popular due to the ease of use. The process involves changing the total number of class data items by either increasing the minority class [17], [18], known as oversampling or reducing the majority class, known as undersampling.

The oversampling techniques were made popular by the pioneering work of [17] through a process called the synthetic minority oversampling technique (SMOTE), which involves artificially generating data items to increase the minority class in the dataset to the level where the imbalance ratio (IR), which is the ratio of the majority class to the minority class, is approximately equal. Although this SMOTE technique apparently has many advantages, particularly in solving the issues of class imbalance, it invariably introduced issues such as the misclassification cost ratio [19], and some researchers have also encountered problems of overfitting, which stem from creating a replica of the same dataset and inheriting the intrinsic errors therein.

Therefore, new approaches are necessary to solve the issues of class imbalance, such as investigating the relationships between variables and the measure of central tendencies, which is our approach. Other modifications of oversampling have been proposed, such as random oversampling used by [20] and [21], which tends to select the training data by random selection. This method, though improving accuracy, has led to delays in the execution and overfitting when dealing with large datasets. A generative oversampling technique was used by [22] and [23], and the process involved new data being created by learning from the training data. This method made it possible for the created data to have the basic characteristics of the existing data, thereby maintaining the data integrity, but the accuracy improvement is limited because the characteristics of the training data are still maintained.

An alternative method, which is the opposite of oversampling is called undersampling, it basically reduces the number of majority classes. These methods have also gained strong research interest in academia. The literature [15] presented two methods of undersampling as random and informative; the random process chooses and eliminates data from the existing class until the class distribution is balanced, while the informative undersampling eliminates the data observation class from the dataset based on preselected criteria to achieve balance. A process known as active undersampling that eliminates the sample of data items that are far from the decision boundary was used by [24]. These sampling methods have a problem with performance in large datasets and can lead to the removal of important data items.

Multiple resampling techniques were employed by [25] because it provides better tuning results with every resampling iteration. A method of integrating the oversampling technique with cross-validation to improve the general performance was proposed by [26]. Cluster sampling methods were also used by [27], which introduced the process of cluster density and boundary density thresholds to determine the cluster and sampling boundary. The literature [28] used a method called a bidirectional sampling based on K-means clustering, which performed very well with data that had too much noise and few samples. Each of these sampling techniques has its benefits and drawbacks, which are very subjective and depend on the context of the application and usage [29].

A technique that could result in an improved performance might not show the same performance when used in different contexts; therefore, more modifications and improvements in the existing sampling techniques have continued to be presented and developed by researchers based on some local properties of the dataset. For instance, some undersampling methods have incorporated the mean of the attribute values as a metric for the derivative of the sampling techniques [30]. One of the main disadvantages of the oversampling method is the risk of overfitting due to generating a replica of the existing data [31]. For undersampling, the main disadvantage is the possibility of discarding some data that might present potential useful information, particularly during the process of variable selection that is cross dependent on other variables or when the potential target is far away from the central data items.

B. ALGORITHM MODIFICATIONS TECHNIQUE

This technique tends to interact with the classification algorithm, making it less sensitive to the class imbalance [32], [33], depending on the derivative of the classification algorithm [34] and proposes a benchmark to validate other algorithm-dependent techniques; it has different nonstandard variations that have been discovered over time; for example, in SVM the margin of class separation is weakened to align the hyperplane to accommodate extraneous classes by adjusting the class boundary in the kernel functions for a condition in which the training data could be represented in a vector space or using a kernel matrix if it is sequential data [35], [36].

A modification of the K-nearest neighbor, called the weighted-K-nearest neighbor (W-KNN) was used by [37]. The process utilized a wider region around the data item distribution to deduce the nearest neighbor, but this has resulted in accommodating some extraneous data, such as outliers, which may add some noise, resulting in the whole prediction becoming less accurate when applied to datasets that has large variances. Recently, a new approach of handling imbalanced datasets known as conditional generative adversarial networks (cGAN) was introduced by [38], which is based on a concept of continuous competitions by two vectors known as generators and discriminators, while the discriminator tries to learn the actual data set pattern by comparing it to the data being generated by the generator, feedback between the

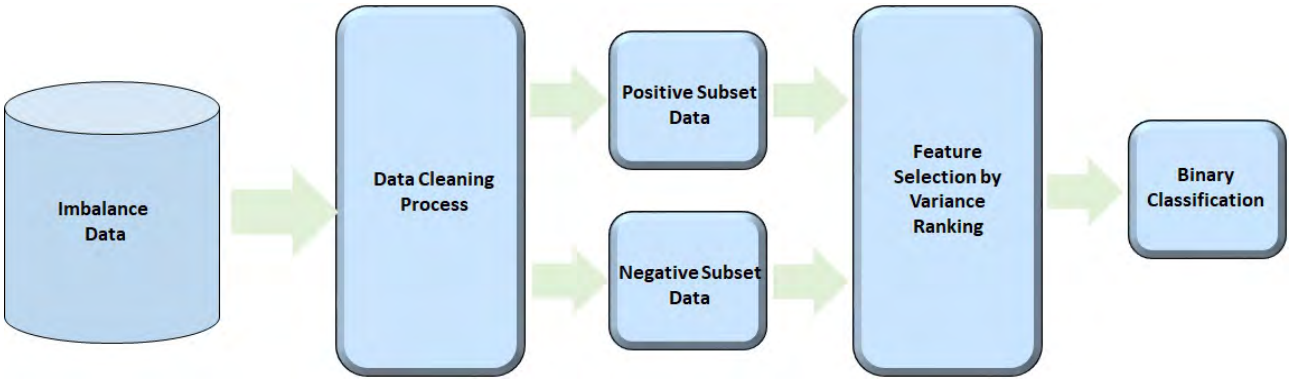

FIGURE 3. An overview of the proposed method.

two vectors results in adaptation and improvement to the data quality.

C. COST SENSITIVE APPROACH

This approach considers the cost of misclassification and adjust the results into empirical consequences of misclassification by allotting a different cost threshold to the target class [39]. In a cost-sensitive approach, the accuracy is not as important as the implication of wrongly classifying the target. In this context, cost might allow a loose boundary, in predicting chronic diseases, an incorrect prediction such as not being able to capture the presence of illnesses or any other unpleasant occurrence; hence, results are computed with a class that leads to minimum cost [40]–[42].

The cost-sensitive learning (CSL) was implemented in combination with resampling and an imbalance ratio by [43]. When using the cost-sensitive approach, each scenario must have a baseline for measuring the acceptable cost and may be modified based on the context of the situation. Combination algorithms such as the ensemble and bagging approaches, although in their infancy, are becoming popular for handling imbalanced datasets. For instance, in [44] and [45], balance-bagging was utilized to study the characteristics of data items as it affects the class distributions.

All the existing methods, as explained in this section, differ from our approaches because they are based on two main methods. Firstly, its either artificially generate or reduce the existing data items to equalize the class distributions. Secondly, the existing machine learning algorithms are modified. However, our approach is based on looking at the intrinsic properties of the attribute distributions within a term of reference (measure of central tendency); we consider how the attribute values are distributed in context to the domain (variable) being measured. We believe that the attribute value arrangement within a central term of reference can provide insight into the relevance of such attributes to the class that they belong to in a binary classed context, and hence, they can be metrics to predictions in binary classification.

III. PROPOSED METHOD AND APPROACH

The classification of imbalanced data poses a significant challenge capable of skewing sensitivities to the majority class target. In addition, the proportions of captured minority target classes may be below the actual numbers of the minority class in the dataset. To achieve proper classification while retaining the sensitivities and features of the data, we propose the technique as illustrated in Figure 3.

By definition, variance is a measure of the spread of the data items of number N, from the mean and is given by $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ if the whole population is considered. However, this equation is slightly different, when a sample of the whole population is used; $\sigma^2 = \frac{\sum(x-\mu)^2}{N-1}$. Consequently, to what extent does the type of variable data item affect the overall variance? Variables are made up of discrete and continuous data items, as in the case of our dataset, and the effect of these intrinsic properties of the data item can be deduced accordingly. For an input dataset with $N = \{n_1.....n_x\}$, where n is a combination of discrete and continuous variables [46]–[48], if the variance of the continuous random variable x is given by Var_x which is the expected value of the square of the the deviation, for all the variable x with a mean of μ the variance is;

$$Var_x = E\left\{(x-\mu)^2\right\} \tag{1}$$

The Equation 1 [47] is modified to accommodate both discrete and continuous variables. Hence, when variable x is continuous the variance is

$$Var_x = E\left\{(x-\mu)^2\right\} = \sum_{i=1}^{n}(x_i-\mu)^2 f(x) \tag{2}$$

Also for discrete variable Equation 1 would resolve to;

$$Var_x = E\left\{(x-\mu)^2\right\} = \int_{-\infty}^{\infty}(x-\mu)^2 f\,(x)\,dx \tag{3}$$

When the whole population is considered, the population variance becomes subjective to the probability density function $f(x)$ such that that the expectation values and mean of x

is the sum of Equation 2 and Equation 3 $V_{(total)} = V_{(disrete)} + V_{(continuous)}$ ie sum of discrete and continuous variable. Therefore, if for independent random variables the variance of their sum or difference is the sum of their variances:

$$V_{total} = \sum_{i=1}^{n} (x_i - \mu)^2 f(x) + \int_{-\infty}^{\infty} (x - \mu)^2 f\,(x)\,dx \quad (4)$$

$$\int x^2 f\,(x)\,dx - \mu^2 \quad (5)$$

if μ is considered as being the propability density functions expected value of x which is equal to $xf\,(x)\,dx = pi$ then the population variable, $V_{(ar)}$ is also deduced by [49]:

$$V_{(ar)} = \sum_{i=1}^{n} pi.\,(x_i - \mu)^2 \quad (6)$$

For all values of pi being the probability density functions, for equation 5 and 6, the similarity is deduced by equating the integral to $f\,(x)\,dx$ and $\sum_{i=1}^{n} pi$. Due to the premise of the same range of probability density function, the variables transformable vis-a-vis discrete and continuous as provided by [46], [50], and [51]. This link between the discrete and continuous distribution under the condition of the same range [52] therefore equalized the variables; hence, the variances of the discrete and continuous variables are equal if $f\,(x)\,dx$ and $\sum_{i=1}^{n} pi$ are equal. Our technique implemented the concept of sample variance by taking n values in the range of $y_1 \ldots . y_n$ of the population where $n < N$. Estimating the variance of the sample data variables gives the average of the square deviations as in $\sigma_2^y = \frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\right) - \bar{y^2} = \frac{1}{n^2}\sum_{i<j}^{1}\left(y_i - y_j\right)^2$. This computation confirms that the range of the variable values of x is still within that of the mean, as explained earlier. This derivative will hold true in both cases of variances if and only if the distribution of the variable x is completely determined by the probability density function $f\,(x)$ [53], [54], which is shown in Equation 5 and 6.

In carrying out the experiments, we have to contend with the properties of the attributes like the continuous and discrete data, the numerous data type and the natural partition in the in the datasets. Therefore, the selection of appropriate statistical techniques that could accommodate these properties on the subpopulation (binary groups). The work of [55], [56] and [57] provided an insight in such processes, the subgroup distribution is not normal as such non parametric statistical techniques as recommended by [58] and [59] will be used. Consequently, we have to considered a process of variance ranking known as Kruskal-Wallis one-way analysis of variance by rank test [60]–[62], which addresses the nonparametric differences between the continuous and discrete variables. This test is used as an alternative to one-way analysis of variance (ANOVA) when a normal distribution in the dataset is not assumed in the probability density functions of $f\,(x)\,dx$ and $\sum_{i=1}^{n} pi$.The Kruskal-Wallis ANOVA by rank is given by

$$H = N - 1\frac{\sum_{i=1}^{g} n_i\,(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_j} n_j\left(\bar{x}_j - \bar{x}\right)^2} \quad (7)$$

Based on the probability density of Equations 5 and 6, the Equation 7 would resolved to

$$\frac{N_{maj}}{N_{min}} = \frac{\sum_{i=1}^{g} n_{maj}\,(x_i - \bar{x})^2}{\sum_{i=1}^{g} n_{min}\left(x_j - \bar{x}\right)^2} \quad (8)$$

The ratio of two random variable events is the same if they share the same probability function and sample space [63], agreeing with Equation 8.

A. VARIANCE SIGNIFICANT TEST

We have to ascertain if the pattern noticed in the variances are significant or just due to mere coincidence, that is the variance of each of the variable in the majority and minority data subsets different from each other? this is done by consideration the **F-distribution** [64]–[66] as against other distributions functions like chi-squared distribution because the F-distribution could deal with multiple sets of events [67] as represented by different variables in the majority and minority data groups or classes. By definition the F-distribution (F-test) [65], [68] is given by

$$F = \frac{Larger\ variance}{smaller\ variance} \quad (9)$$

For the the sum of discrete and continuous variable, will be $F_i = F_{discrete} + F_{continuous}$ as provided in Equation 4,

$$F_i = \left[\frac{V_{1d}}{V_{2d}} + \frac{V_{1c}}{V_{2c}}\right] \quad (10)$$

Therefore, for subset (class 1 and 0)

$$F_{final} = F_i + F_j \quad (11)$$

The Equation 10 and 11 agreed with the rules that ''variance of independent variable is additive'' please see [69], [70] Var[X+Y] = Var[X]+Var[Y].

The unit of F_{final} is the same unit as variance, hence F_{final} is a measure of variance of the variances F_1 *and* F_j. For a Binary classed data set, if the sub classes variance is V_i and V_j, then the 11 would resolved to 12, the squaring is a mathematical expediency to eliminate any negative in the value of F_{final}

$$F_{final} = \left[\frac{V_{1i}}{V_{0i}} + \frac{V_{0j}}{V_{1j}}\right]^2 \quad (12)$$

B. SPLITTING THE DATA SET

In ***splitting*** of the data set, the total number of instances were taken into consideration, were the numbers are below a thousand like in (Pima, Wiscosin and Bupa) all the data were taken and split into two (training and test) data by the ratio of 60% and 40%, the reason for taken this close proportion is to avoid ending up with very few numbers of minority groups in both the training or the test data since it will be split again into positive and negative down along the line. For Cod-rna data set random selection of one 1000 instance from the total of 488565 in the proportion of 67% and 33% (see Table 1) which is the ratio of the positive and negative instances in the

TABLE 1. Table of the four data set.

		Pima Indians Diabetes data	Wisconsin breast cancer	BUPA liver disorders	Cod-rna Dataset
Attributes + class		8	11	7	9
Number of Instances		768	699	345	488565
Negative		500 (65.10%) =0	Benign: 458 (65.5%) =0	200 (57.97%) =0	325,710 (67%) =0
Positive		268 (34.90%) = 1	Malignant:241 (34.5%) =1	145 (42.20%) =1	162,855 (33%) =1
Missing Value		yes	none	yes	none
Number of Class		2	2	2	2

total data set, this was done before splitting it further to 60% and 40% of (training and test) data set.

The training data set was divided into two subsets of different classes to represent the two events (positive and negative). If the variance of subset V_0 (Variance of negative class) is $\sum_{i=1}^{g} n_{maj} (x_i - \bar{x})^2$ and V_1 (Variance of positive class) is $\sum_{i=1}^{g} n_{min} \left(x_j - \bar{x}\right)^2$ of the data set were obtained, the ratio of V_0 to V_1 was added to the ratio of V_1 to V_0, and we squared the results to eliminate any negative values. The results were then ranked to achieve the final classification. The criteria used for evaluation are as follows. First, we compared the results obtained with two benchmarks of attribute selection (Pearson correlations and information gain), and in the second evaluations, we provided a series of experiments of binary classification: logistic regression (LR), support vector machine (SVM) and decision trees (DT) without attribute selection and with attribute selection obtained in our variance ranking techniques. We describe in detail the processes in Section 4.

C. DATA ACQUISITION AND DESCRIPTIONS

The datasets used in this research are the Wisconsin Breast Cancer data, the BUPA liver disorders data, the Pima Indians Diabetes data and the Cod-rna data. The datasets are from the machine learning archive [71], [72], the full descriptions and other details of the datasets are listed in Tables 1.

The data are in the public domain; hence, no extra permission was needed. All references to the data have been acknowledged. Detailed descriptions such as the number of instances, total attributes, missing data, and class distributions are all provided. The main similarities in the datasets are that the target classes are all binary (two classes).

D. DATA PREPARATION

Although the Weka data mining and machine learning software were used for most analyses, we also used Microsoft Excel for initial analysis and data preparation, such as counting the missing values and descriptive statistics. The work involved four datasets, Pima Indians Diabetes data, Wisconsin Breast Cancer data, BUPAliver disorders data and the cod-rna dataset (please see Section III-C). These datasets were explored to ascertain the types of data preparations that would be applied to each in accordance with Cross Industrial Standard Process- CRISP for data mining [73], [74].

For the ***Missing data***, two of the datasets had missing data; the Pima Indians Diabetes data and the BUPA Liver Disorders data. This was treated using the average of the data column items because the Skewness for the missing columns are zero, hence their mean value were used as replacement for the missing data in the body mass index (BMI) and age attributes in the Pima Indians Diabetes data, also for the BUPA Liver Disorders data, the aspartate aminotransferase (sgot) and alanine aminotransferase (sgpt) columns were also treated for missing data values. The Wisconsin Breast Cancer data are well organized and were treated from source, so there were no problems with the data, while the cod-rna dataset had very few cases of missing values; thus, it was deleted. Additionally, none of the data had any problem with outliers. Finally, the inconsistency of representing missing values with zero in the Pima Indians Diabetes data was also addressed in the BMI column. For any part of this research, three binary classification algorithms (DT, LR and SVM). These three Machine Learning algorithm belong to the group of supervised learning algorithms, DT have the advantages of being able to handle both regression and classification problems, in its simplest form it takes the training data set as the root node and split it into two and continue splitting until the final node, see Figure 4.

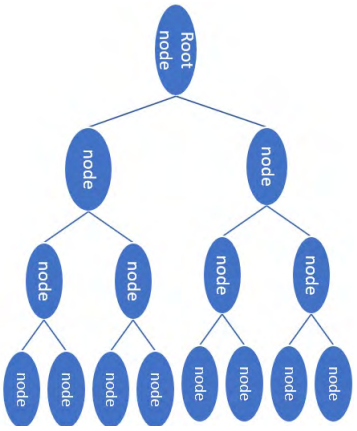

FIGURE 4. Decision tree.

The splitting is based on intrinsic properties of the data set known as Entropy *Entropy* $H(X) = -\sum p(X) \log p(X)$ which is a measure of the homogeneity of the data therein and is derived from either information Gain *Information Gain* $I(X, Y) = H(X) - H(X|Y)$ or Gini index $Gini(E) = 1 - \sum_{j=1}^{n} (p)^2$ [75]–[77].

The LR is an Logistic Regression is an offshoot of linear regression $y = mx + c$, which is basically an equation of straight line. LR uses a logistic sigmoid function to transform its output into a probability discrete binary like(1 or 0, yes or no, true or false) see Figure 5, LR is very easy to implement, interpret and train [78]–[80].

The SVM algorithm considered all data items as a point in a plane where a dividing line that demarcate (the hyperplane) the data points into two parts representing the binary classes which the data belong to see Figure 6 .

Any new test data will either belong to one of the classes. Apart from some of the advantages of these three

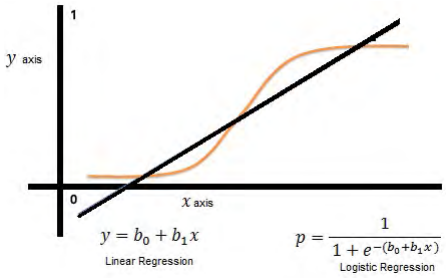

FIGURE 5. Logistic regression.

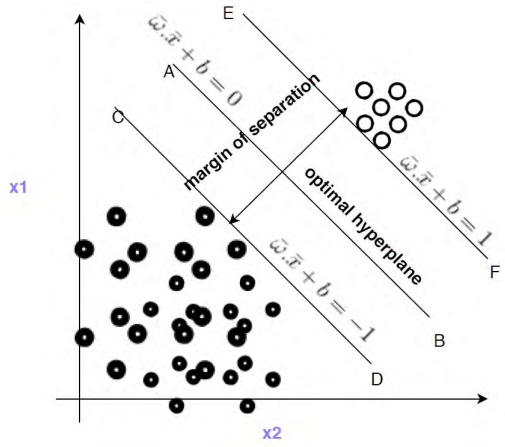

FIGURE 6. Support vector machine.

algorithms mentioned earlier, the main reasons for selecting these algorithms is because of their performance with correlated attributes that are not directly related to the outputs but could be related to other attributes, these machine learning algorithms could detect such indirect relations [81]. We also used k-fold cross-validation to avoid over-fitting. The metric of measurement in the binary context is addressed in the next section.

E. RESULT AND MATRIX TERMS DEFINITIONS

While there are different binary classification algorithms, all the performance evaluations uses the confusion matrix [82]. The confusion matrix is represented by a table summary of numbers in a group that have been correctly and incorrectly predicted [83], [84] therefore, the classification results will be explained using a confusion matrix, which is a cross-section table that evaluates how accurate the model classifies the binary groups. One major reason for using this metric in measuring binary classification is the insight into how the algorithm identifies the classes and how many classes have been confused and mislabeled and the ability to visualized the classification performance as explained in [85] and [86]. This enables the assessment of the accuracy of the model to be easily compared to the benchmark.

The confusion matrix in Table 2 is especially useful in binary situations as against multiclass classification, where

TABLE 2. Confusion matrix.

	Predicted	
	Positive	Negative
Actual Yes	TP	FN
Actual No	FP	TN

multiple overlapping classifications could make the result less discriminant. The terms in the confusion matrix tables are defined below:

True positives (TP): *These are cases in which the system predicted yes and they do have the disease.(Correctly predicted).*

True negatives (TN): *We predicted no, and they don't have the disease.(Correctly predicted).*

False positives (FP): *We predicted yes, but they don't have the disease.(Incorrectly predicted)*

False negatives (FN): *We predicted no, but they do have the disease.(Incorrectly predicted).*

True Positive Rate (TPR) = Sensitivity = Recall; defined as the proportion of actual positives which are predicted positive).

$$Recall = \frac{TP}{(TP + FN)}. \tag{13}$$

$$Precision = \frac{TP}{(TP + FP)}. \tag{14}$$

$$F = \frac{Precision.Recall}{(Precision + Recall)}. \tag{15}$$

$$Accuracy = \frac{tp + tn}{(tp + tn + fp + fn)} = \frac{tp + tn}{n}. \tag{16}$$

The formulas show that the F-score is another means of testing the accuracy of the binary classification [87].

IV. EXPERIMENT

A. VARIANCE RANKING FEATURE SELECTION METHOD

The proposed method of attribute selection should depend on the nature of the data and the target. For instance, in discrete and continuous numeric data with a binary class target of 1 or 0, variance ranking has been found to show a very promising result. Missing values and other data preparations: the first step was to treat the missing values and other necessary data preparations as described above in the Pima Indians Diabetes data, Wisconsin Breast Cancer data, BUPA Liver Disorders data and the cod-rna data, which have a high imbalance ratio. Note that none of the datasets has the problems of outliers. The missing values were treated by replacing them with the average of the data in each of the columns that had missing values. The variables are numeric with a binary target (0, 1 or as detailed in the dataset). The aim was to find which of these variables could be very significant in predicting the target class. First, a subsection of the dataset that belongs to each target class was selected, e.g., 1 and 0. The variance of each of the subsections, class target 1 and class target 0, of the dataset was computed using the following variance formula $v = \frac{\sum(x-\bar{x}^2)}{(n-1)}$. If The Variance subsection of class 0 is

given by:

$$V_0 = \frac{\sum_{i=1}^{n}(x_0 - \bar{x_0}^2)}{(n-1)} \tag{17}$$

If then Variance subsection of class 1 is given by:

$$V_1 = \frac{\sum_{i=1}^{n}(x_1 - \bar{x_1}^2)}{(n-1)} \tag{18}$$

The Variance Ranking is then deduced by:

$$VR = \left(\frac{V_1^2 + V_0^2}{V_1 V_0}\right)^2 \tag{19}$$

being a derivatives from equations 9 to 11 as applied to both the majority and minority classes in the data set.

TABLE 3. The variance ranking of Pima India Diabetes data.

Variable	V_1	V_0	(V_1/V_0)	(V_0/V_1)	VR
plaglu	830.318	683.028	1.2156427	0.822610133	4.154474597
bmass	52.5539	42.9658	1.2231566	0.817556832	4.164511137
age	119.856	116.344	1.0301863	0.970698171	4.003538843
preg	10.1388	7.67682	1.3207031	0.757172446	4.317566802
insutest	10583.2	8073.62	1.310837	0.762871343	4.300266385
skinfold	98.4056	70.0424	1.4049433	0.711772501	4.480485745
pedi	0.13812	0.09846	1.4028032	0.712858384	4.476023806

TABLE 4. The variance ranking of liver disorder Bupa data.

Variable	V_1	V_0	(V_1/V_0)	(V_0/V_1)	**VR**
mcv	14.96964	23.08621	0.6484235	1.5422021	4.798840241
alkphos	345.6176	326.2356	1.0594111	0.9439207	4.013338026
sgpt	248.943	477.2004	0.5216739	1.9169063	5.946673357
sgot	59.87759	127.4371	0.46986	2.1282937	6.750402623
gammagt	1103.902	1807.82	0.6106261	1.6376635	5.054805883
drinks	15.44272	8.075069	1.9123948	0.5229046	5.930683086

TABLE 5. The variance ranking of Wisconsin breast cancer data.

Variable	$\mathbf{V_1}$	$\mathbf{V_0}$	$(\mathbf{V_1/V_0})$	$(\mathbf{V_0/V_1})$	**VR**
ClumpThickness	5.8993084	2.8033406	2.1043852	0.4751982	6.654250436
Uniformity of Cell Size	7.3957469	0.8239085	8.9764174	0.111403	82.58847934
UniformityofCellShape	6.5640733	0.9956762	6.5925784	0.1516857	45.48509828
MarginalAdhesion	10.307089	0.9936696	10.372753	0.0964064	109.6032918
SingleEpithelialCellSize	6.0103734	0.8411273	7.1456165	0.1399459	53.07942022
Bare Nuclei	9.713688	1.3873264	7.0017324	0.1428218	51.04465399
BlandChromatin	5.1704011	1.1671333	4.4300005	0.2257336	21.67586052
NormalNucleoli	11.227006	1.1211767	10.013592	0.0998643	102.2819975
Mitoses	6.5430498	0.2519995	25.964538	0.0385141	676.1587396

TABLE 6. The variance ranking of cod-rna data.

Variable	V1	V0	(V1/V0)	(V0/V1)	VR
X1	1.002028	0.9863	1.015946	0.984304	4.001001
X2	0.939372	0.955021	0.983614	1.016659	4.001092
X3	0.98205	0.998551	0.983475	1.016803	4.001111
X4	0.992204	1.016567	0.976034	1.024554	4.002354
X6	0.999488	1.025702	0.974443	1.026227	4.002682
X5	0.994258	0.959546	1.036175	0.965088	4.005053
X8	0.999864	1.042645	0.958969	1.042787	4.007025
X7	0.950464	0.642895	1.478412	0.676401	4.643222

In Tables 3, 4, 5 and 6, the column V_1 and V_0 are the results of the variance of each subsection class (positive=1 and negative=0) for each attribute. The column (V_1/V_0) and (V_0/V_1) is the division of each variance and inverted and divided again to produce the values that could be added to each other. The $VR = \left(\frac{V_1+V_0}{V_1V_0}\right)^2$ is the addition of the two columns (V_1/V_0) and (V_0/V_1) and finally, the result is squared.

B. COMPARISON OF VARIANCE RANKING FEATURE SELECTION WITH OTHER BENCHMARKS TECHNIQUE

Attribute selections in general could be categorized as filter and wrapper methods [88], [89]. The filter method uses the general characteristics of the data item to determine the features that are more significant without involving any intended learning algorithm, while wrapper method on the order hand tend to determined the features in data set that would produce the best performance on a predetermined learning algorithm, putting it succinctly, wrapper method suggest the attributes to use for a given classifier. This suggestive and predetermining the classifier made the wrapper method less generic and limited as a means of comparison with our method (Variance raking) which is independent of any learning algorithm. Beside wrapper methods create a subset of features which are deemed to be most importance for a specific classifier's performance, these subsets more often than not does not contained all the original features meaning that some features are eliminated in the subsets and each feature relevance to the subsets are not made known, but filter methods uses all the features and rank them to produce the order of relevance of each, ie no feature is eliminated from the ranking [88] The comparison of variance ranking attribute selection will be done with similar filter method that are not classifier suggestive and uses all the features. Consequently, We compare our method to the state-of-art filter feature selection methods; the Pearson correlation (PC) and [90]–[92] and information gain (IG) [93], [94] the results are provided in Tables 7, 8, 9 and 10 for the four data sets used in the experiment.

Variance(V) is given by:

$$V = \frac{\sum(x - \bar{x}^2)}{(n-1)}. \tag{20}$$

Pearson Correlation(PC) is given by:

$$PC = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}. \tag{21}$$

and Information Gain$_{(x,y)}(IG)$ between X and Y is

$$IG = Entropy_{(x)} - Entropy_{(x,y)}. \tag{22}$$

TABLE 7. Comparison of variance ranking with PC and IG variable selection for Pima India Diabetes data.

Ranking of variables based on different feature selection Algorithm			
sn	**Variance Ranking**	**Pearson Correlation**	**Information Gain**
1	age	plasglu	Age
2	plasglu	bmass	bmass
3	bmass	preg	plasglu
4	insutest	age	preg
5	preg	insutest	insutest
6	pedi	skinfold	skinfold
7	skinfold	pedi	pedi
8	diapres	diapres	diapres

Tables 7, 8, 9 and 10 show the results obtained using the three attribute selections on the four binary classed datasets; Pima Indians Diabetes, BUPA Liver Disorders, Wisconsin Breast Cancer data and the cod-rna data ranked the attributes according to their relevance to the target class (1, 0). The four

TABLE 8. Comparison of variance ranking with PC and IG variable selection for liver disorder Bupa data.

Ranking of variables based on different feature selection Algorithm			
sn	Variance Ranking	Pearson Correlation	Information Gain
1	sgot	sgot	drinks
2	sgpt	gammagt	gammagt
3	drinks	drinks	sgpt
4	gammagt	mcv	sgot
5	mcv	sgpt	alkphos
6	alkphos	alkphos	mcv

TABLE 9. Comparison of variance ranking with PC and IG variable selection for Wisconsin breast cancer data.

Ranking of variables based on different feature selection Algorithm			
sn	Variance Ranking	Pearson Correlation	Information Gain
1	ClumpThickness	UniformityofCellShape	Uniformity of Cell Size
2	BlandChromatin	Uniformity of Cell Size	BlandChromatin
3	UniformityofCellShape	Bare Nuclei	UniformityofCellShape
4	Bare Nuclei	BlandChromatin	Bare Nuclei
5	SingleEpithelialCellSize	ClumpThickness	SingleEpithelialCellSize
6	Uniformity of Cell Size	NormalNucleoli	NormalNucleoli
7	NormalNucleoli	MarginalAdhesion	ClumpThickness
8	MarginalAdhesion	SingleEpithelialCellSize	MarginalAdhesion
9	Mitoses	Mitoses	Mitoses

TABLE 10. Comparison of variance ranking with PC and IG variable selection for cod-rna data.

Ranking of variables based on different feature selection Algorithm			
sn	Variance Ranking	Pearson Correlation	Information Gain
1	X1	X1	X1
2	X2	X2	X2
3	X3	X4	X3
4	X4	X5	X4
5	X6	X8	X5
6	X5	X7	X7
7	X8	X3	X6
8	X7	X6	X8

results were comparatively similar but had some minor differences. For instance, in Table 7, the most significant attribute using variance ranking and information gain was (age) for the Pima Indians Diabetes data, while the first in the Pearson correlation was plasma glucose; in row numbers 6, 7 and 8 of Table 7, the three attribute selection techniques selected pedi, skinfold and diapres, respectively, as the least significant attribute in a slightly different order. Also in the BUPA Liver Disorders data in Table 8,the most significant attribute using variance ranking and the Pearson correlation was agot while sgot ranked third by the information gain selection, but in Table 8, in row numbers 5 and 6, each of the attribute selection techniques selected mcv and alkphose, respectively, as the least significant attributes. For Table 9 For the Wisconsin Breast cancer data, two of the techniques (variance ranking) and (information gain) were in agreement selecting Mitoses and MarginalAdhesion as the least significant attributes, while the Pearson correlation also identified Mitoses as the least significant attribute but selected SingleEpithelialCellSize as the second least significant attribute. For Table 10 for the cod-rna data, the variance ranking and the information gain techniques were similar in rows 1, 2, 3 and differed slightly in rows 4 and 5. However, clear similarities were very noticeable for all three techniques. Generally, the three selection methods identified the same sets of attributes but ranked them in a slightly different order.

C. COMPARISON OF VARIANCE RANKING FEATURE SELECTION TO OTHERS USING RANKED ORDER SIMILARITY-ROS

Similarity and dissimilarity measure has been used to to compare item and results of two or more structures, but quite recently many data centric researches like data mining and machine learning have used this process to compare and validate the results [95]–[97] of experiments and predictive modeling, this is done by measuring the similarity index of a new concept with existing benchmarks knowledge, concept or results. With this in mind we proposed a novel similarity measure technique-ROS, we want to determine how similar the results are in Tables 7, 8, 9 and 10, Considering Table 7 for instance. Should we say that the result of variance ranking and the Pearson correlation are 50%,70%, 80% or 90% similar? How should their similarities be graded? Although there are different approaches to measuring the similarities, the five most popular are Euclidean distance, Manhattan distance, Minkowski distance, cosine similarity and Jaccard similarity [98], [99]; however, none of these is appropriate to measure the similarities between two or more sets that contain the same objects but are arranged or ranked differently. If three Sets $\alpha = \{a, b, c, d, e, f\}$, $\beta = \{a, b, c, f, e, d\}$ and $\gamma = \{f, b, c, d, e, a\}$ contain the same elements arranged or ranked in a different order as in Table 11, based on the order of ranking, what are the percentage similarities?

TABLE 11. Three sets arranged and ranked in different order.

sn	α		β		γ
1	a		a		f
2	b		b		b
3	c		c		c
4	d		f		d
5	e		e		a
6	f		d		e

Let us determine the similarities between α and β. The total elements in α and β is 12 ie $N = 12$, Since we wish to find the percentage similarities, we use Equation 23 we defined a quantity which is called *Element Percentage Weighting* = *EPW* given by:

$$EPW = \sum \frac{100}{N} \tag{23}$$

Therefore, $100/N = 8.33$; thus,each element of the set has a percentage weighting of 8.33%; two elements in a row would have a total percentage weighting of the sum of their weightings; for example, in row 1 in Figure 7, the total percentage weighting of an element a in Set α and element a in Set β is $8.33 + 8.33 = 16.66$.

Additionally, each set has a total number of n. When an element moves downward or upward in a column to be in the same row with its similar element, it loses a percentage weighting equal to $EPW - (2 * \frac{EPW_j}{n})$. the 2 is because there are two elements. The quantity $(\frac{EPW_j}{n})$ is called the unit *Element Percentage Weighting*. The sum of all the $(\frac{EPW_j}{n})$ is equal to the *EPW* for the two set:

$$ROS = \sum_{1-J}^{n} EPW - \sum_{1-j}^{n} 2 * \frac{EPW}{n} \tag{24}$$

Calculation of ROS between α and β sets					
sn	α	β	ΣEPW	ΣEPW	Comments
1	a	a	8.33+8.33	16.66	value is ΣEPW and note (EPW/n) =0), because the elements did not move
2	b	b	8.33+8.33	16.66	value is ΣEPW and note (EPW/n) =0), because the elements did not move
3	c	c	8.33+8.33	16.66	value is ΣEPW and note (EPW/n) =0), because the elements did not move
4	d	f	4.166 + 4.166	8.33	(2*EPW/n) f moved 3 position (3*1.388 = 4.164) 8.33 - 4.164 = 4.166 2*4.166 = 8.33
5	e	e	8.33+8.33	16.66	value is ΣEPW and note (EPW/n) =0), because the elements did not move
6	f	d	4.166 + 4.166	8.33	(2*EPW/n) d moved 3 position (3*1.388 = 4.164) 8.33 - 4.164 = 4.166 2*4.166 = 8.33
if EPW =	100/12	= 8.33	$ROS = \sum_{1-j}^{n} EPW - \sum_{1-j}^{n} 2 * \frac{EPW}{n}$		**=83.3%**
if (EPW/n)	8.33/6	=1.388			

FIGURE 7. Ranked order similarity-ROS percentage weighting calculation for α and β.

In rows 4 and 6 for sets α and β, the element d and f are not in the same row with their similar item. To calculate their weighting using Equation 23 is given by 8.33-$\sum$ *Loss percentage weighting*, if *Loss percentage weighting* $= EPW/n = 8.33/6 = 1.388$. Elements d and f have moved up and down three steps (including their row), the total Loss percentage weighting for each is $3 * 1.388 = 4.164$, and the final weighting for each is $8.33 - 4.164 = 4.166$. Therefor in row 4, $f + f = 4.166 + 4.166 = 8.33$. Additionally, Also in row 6, $d + d = 4.166 + 4.166 = 8.33$. The similarity between sets α and β is 83.3%, Please see 24 and Figure 7 represents the process of calculating the ranked order similarity-ROS.

TABLE 12. Comparison of varinace ranking, Pearson correlation and information gain using ROS.

Pima India diabetes			
	Varinace Ranking	Pearson Correlation	Information Gain
Varinace Ranking	100	74	81.25
Pearson Correlation	74	100	86
Information Gain	81.25	86	100
Bupa Liver Disorder data			
	Varinace Ranking	Pearson Correlation	Information Gain
Varinace Ranking	100	75	56
Pearson Correlation	75	100	58.35
Information Gain	56	58.35	100
Wisconsin Breast cancer			
	Varinace Ranking	Pearson Correlation	Information Gain
Varinace Ranking	100	68	82
Pearson Correlation	68	100	78
Information Gain	82	78	100
Cod-rna data			
	Varinace Ranking	Pearson Correlation	Information Gain
Varinace Ranking	100	69	84.38
Pearson Correlation	69	100	77
Information Gain	84.38	77	100

Table 12 is the comparison table using the ranked order similarity-ROS to compare the results of variance ranking, Pearson correlation and information gain attributes selection techniques in Tables 7, 8, 9 and 10. In the Pima Indians Diabetes database, variance ranking and information gain are 81.25% similar, while it is 74% similar to Pearson correlation. Even the Pearson correlation is 86%, which is similar to the information gain. In the BUPA Liver Disorders data, variance ranking is 75% similar to the Pearson correlation, while it is 56% similar to the information gain, and the Pearson correlation is 58.35% similar to the information gain. In the Wisconsin Breast Cancer data, variance ranking is 68% similar to the Pearson correlation and 82% to the information gain, while the information gain and Pearson correlation are 78% similar. Finally, in the cod-rna data variance ranking is 84.38% similar to the information gain and 69% similar to the Pearson correlation, the Pearson correlation and the information gain are 77% similar. This comparison establishes the following facts: (1) the efficacy of the ROS similarity measure and (2) no two attribute selection processes produce 100% of the same results.

D. VALIDATION OF VARIANCE RANKING ATTRIBUTES SELECTION USING BINARY CLASSIFICATION

In this section, the variance ranking variable selection will be tested; it involves carrying out binary classification using the following three algorithms: logistic regression (LR), support vector machine (SVM) and decision tree (DT). Three of the datasets described in Tables 1 (Pima Indians Diabetes, Wisconsin Breast Cancer data and Bupa liver disease data) will be used to make the following predictions: (1) Who among the patients is likely to be diabetic in Pima Indians Diabetes data? (2) Which of the tumors are malignant in the Wisconsin Breast Cancer data? and who is most likely to have liver disease from the Bupa data set The two target classes are 0 or 1. A confusion matrix is used to deduce the accuracy of the binary classifications, and the details and explanations have been provided in Section III-E (Results and matrix terms definitions). The following will be deduced from the confusion matrix: true positive (TP), true negative (TN), false positive (FP) and false negative, the $F_{Measure}$ and the receiver operating characteristics (ROC).

1) THE SELECTION OF THE ATTRIBUTES FROM PEAK ACCURACY THRESHOLD

The attributes selection has been carried out and compared with the benchmarks in section IV-A and using ROS to quantify the similarities of variance ranking and others benchmarks in sections IV-B. The attributes have been ranked from the most to the least significant, the next stage is how to use this ranking; meaning which attributes should be selected or eliminated as the case maybe? the work of [100]–[102] provided an exposition by reevaluating the concept of "Most and Least" significant attributes, taking a cue from this concept, we made the following postulations; For any data set A with attributes $a_1 \ldots . a_n$ ranked from the "Most" to the "Least" significant, if we start with the most significant from the rank $a_1 \ldots . a_n$ and continued adding the attributes toward the "least" significant, the accuracy (or the capture of the $TPRate_{min}$) of the prediction would increase and

TABLE 13. Experiment without attribute selection for LR, SVM and DT for three data set.

	Data Set	Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	% Accuracy
Without Attribute Selection	Pima	L R	0	0.88	0.429	0.793	0.88	0.834	0.832	
			1	0.571	0.12	0.718	0.571	0.636	0.832	77.22%
		SVM	0	0.898	0.459	0.785	0.898	0.838	0.72	
			1	0.541	0.102	0.74	0.541	0.625	0.72	77.34%
		DT	0	0.766	0.437	0.766	0.766	0.766	0.709	
			1	0.563	0.234	0.563	0.563	0.563	0.709	69.53%
	Wisconsin	L R	0	0.954	0.137	0.93	0.954	0.942	0.914	
			1	0.863	0.046	0.908	0.863	0.885	0.932	92.27%
		SVM	0	0.965	0.05	0.974	0.965	0.969	0.946	
			1	0.95	0.035	0.935	0.95	0.942	0.958	95.99%
		DT	0	0.952	0.154	0.922	0.952	0.937	0.964	
			1	0.846	0.048	0.903	0.846	0.874	0.965	91.56%
	Bupa	L R	0	0.51	0.435	0.46	0.51	0.484	0.533	
			1	0.565	0.49	0.614	0.565	0.589	0.533	54.20%
		SVM	0	0.51	0.355	0.51	0.51	0.51	0.578	
			1	0.645	0.49	0.645	0.645	0.645	0.578	58.84%
		DT	0	0.497	0.36	0.5	0.497	0.498	0.566	
			1	0.64	0.503	0.637	0.64	0.638	0.566	57.97%

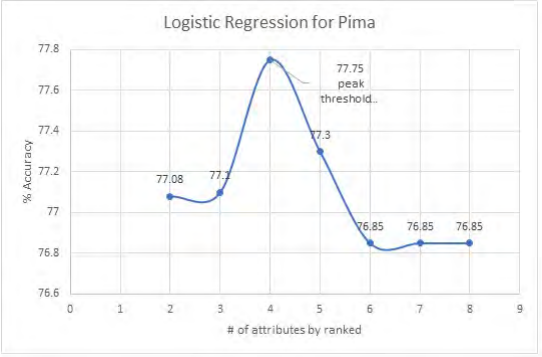

FIGURE 8. Attribute selection by peak threshold accuracy for LR algorithm.

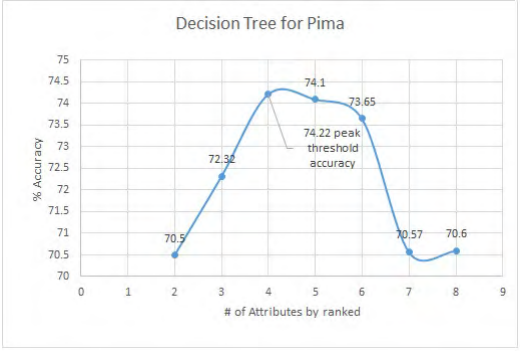

FIGURE 10. Attribute selection by peak threshold accuracy for DT Algorithm.

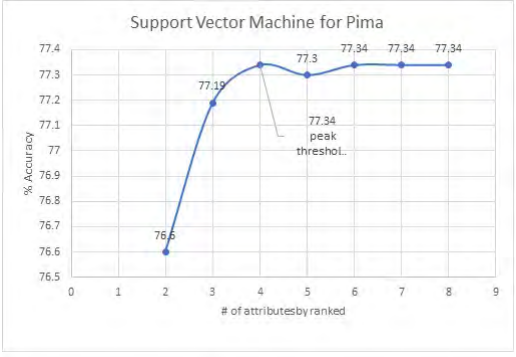

FIGURE 9. Attribute selection by peak threshold accuracy for SVM algorithm.

peak at a threshold of the most significant attributes before decreasing in the accuracy (or the capture of the $TPRate_{min}$). Figure 8, 9 and 10 is a representation of the of the increase of accuracy as the most significant attributes are added until the peak threshold is reached before decreasing or falling in accuracy.

Although, accuracy was used, but the same relationship also exist between the number of the attributes and the $TPRate_{min}$ in Tables (14, 15 and 16) The higher accuracy the higher the captured $TPRate_{min}$.

This peak threshold accuracy techniques were used in all the three data set for selecting the attributes, the important thing here is to note the number of attributes required for our techniques (variance ranking) needed to attained the peak threshold (only 4 attributes) as against total of 8 attributes in the Pima data set. Also using the PC and IG attributes selection it takes a total of 6 attributes to attained the peak accuracy threshold. Therefore not only that variance ranking attributes selection is superior in performance, it is also superior by using fewer attributes to attained higher accuracy.

2) VALIDATION EXPERIMENT OF BINARY CLASSIFICATION USING PIMA INDIA DIABETES DATA

The corresponding experimental results are in Tables (13, 14, 15 and 16) which contains the tabular results obtained using the Pima Indians Diabetes, Wisconsin

TABLE 14. Experiment with attribute selection using VR for LR, SVM and DT for three data set.

	Data Set	Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	% Accuracy
With Attribute Selection Variance Ranking	Pima	L R	0	0.884	0.422	0.796	0.884	0.838	0.825	77.73%
			1	0.578	0.116	0.728	0.578	0.644	0.825	
		SVM	0	0.898	0.459	0.785	0.898	0.838	0.72	77.34%
			1	0.541	0.102	0.74	0.541	0.625	0.72	
		DT	0	0.8	0.366	0.803	0.8	0.802	0.773	74.22%
			1	0.634	0.2	0.63	0.634	0.632	0.773	
	Wisconsin	L R	0	0.976	0.046	0.976	0.976	0.976	0.993	96.85%
			1	0.954	0.024	0.954	0.954	0.954	0.993	
		SVM	0	0.974	0.037	0.98	0.974	0.977	0.968	97.00%
			1	0.963	0.026	0.951	0.963	0.957	0.968	
		DT	0	0.956	0.075	0.961	0.956	0.958	0.955	94.56%
			1	0.925	0.044	0.918	0.925	0.921	0.955	
	Bupa	L R	0	0.531	0.21	0.647	0.531	0.583	0.718	68.12%
			1	0.79	0.469	0.699	0.79	0.742	0.718	
		SVM	0	0.455	0.11	0.75	0.455	0.567	0.673	70.72%
			1	0.89	0.545	0.693	0.89	0.779	0.673	
		DT	0	0.531	0.2	0.658	0.531	0.588	0.665	68.70%
			1	0.8	0.469	0.702	0.8	0.748	0.665	

Breast Cancer and the BUPA Liver Disorders data sets for the following algorithms LR, SVM, and DT, these tables contain the following correctly and incorrectly classified: percentage accuracy, TP rate, FP rate, precision, recall, F-measure and ROC area for both classes (0 and 1), these were deduces from Equations 13, 14, 15 and 16 and the confusion matrix in Table 2. The blow-by-blow details of the experimental results will be provided in the successive sessions. In Tables (13, 14, 15 and 16), the results of logistic regression- LR (with and without) attribute selection, the results of the support vector machine-SVM (with and without) attribute selection and the decision tree-DT (with and without) attribute selection. Notice that in the above tables, class 0 are patients without diabetes (negative) and class 1 are patients with diabetes (positive); the total number of instances is 768, with 500 belonging to class 0 and 268 belonging to class 1; hence, the minority is 268 in number.

The main aim of the above experiment is to show that variance ranking attribute selection improved the sensitivities of the algorithm to capture or target class 1 patients with diabetes. This is shown by the increase in the TP rate for class 1, as indicated in Table 14. From the analysis of the results of LR, the TP rate for without attribute selection is 0.571 and that for attribute selection is 0.578. This is an approximately 1.2% increase in the accuracy of targeting the minority class. The results of SVM also did not show any increase from 0.541 to 0.541 but uses fewer attribute. Finally, the results of DT TP Rate for minority class (1) increases from 0.563 to 0.634, is the biggest increase. This accounted for the additional identification of more 25 patients from the minority of 268.

The results clearly demonstrated unequivocally that the variance ranking attribute selection works and has a direct impact on the general accuracy of the classification model to target the minority in an imbalanced data set.

3) VALIDATION EXPERIMENT OF BINARY CLASSIFICATION USING WISCONSIN BREAST CANCER DATA

In Tables (13, 14, 15 and 16) is the binary classification using LR, SVM and DT on the Wisconsin Breast Cancer data for the predictive model. Firstly, all 9 attributes were used for the predictions (without attributes selections) as in Tables(13) The following the techniques as explained in sections IV-D.1 until a peak threshold with highest accuracy and highest $TPRate_{minority}$ which is class(1). Figures 8, 9 and 10 showed a similar graph of the peak threshold at which the attributes selection was made. Any future removal or addition of attribute(s) after this threshold resulted in the reversal of the general accuracy and specific accuracy of the $TPRate_{minority}$, the results of the peak threshold agreed with the results obtained from the three attributes selection techniques used (variance ranking, Pearson correlations and information gain). The analysis of predictive model for Wisconsin data set in Tables (13, 14, 15 and 16) contain the following results logistic regression (LR), support vector machine (SVM), and decision tree (DT), with and without attributes selections.

For each of the predictions, a summary statistic is provided showing the general percentages accuracy true positive rate (TPRate), true negative rate (TNRate), false positive (FPRate), false negative (FNRate), F-measure and receiver operating characteristics (ROC) for both the majority class 0 and Minority class 1, all these were deduced from the confusion matrix as provided and explained in Table 2. In the LR results, the general accuracy increases from 92.27% to 96.85% with a corresponding increase in $TPRate_{minority}$ class 1 from 0.863 to 0.954, the SVM accuracy increases from 95.99% to 97% and the $TPRate_{minority}$ class 1 increases

TABLE 15. Experiment with attribute selection using PC for LR, SVM and DT for three data set.

	Data Set	Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	% Accuracy
With Attribute Selection by Pearson C	Pima	L R	0	0.88	0.429	0.793	0.88	0.834	0.832	
			1	0.571	0.12	0.718	0.571	0.636	0.832	77.21%
		SVM	0	0.898	0.466	0.782	0.898	0.836	0.716	
			1	0.534	0.102	0.737	0.534	0.619	0.716	77.08%
		DT	0	0.852	0.444	0.782	0.852	0.815	0.791	
			1	0.556	0.148	0.668	0.556	0.607	0.791	74.87%
	Wisconsin	L R	0	0.974	0.05	0.974	0.974	0.974	0.993	
			1	0.95	0.026	0.95	0.95	0.95	0.993	96.57%
		SVM	0	0.974	0.041	0.978	0.974	0.976	0.966	
			1	0.959	0.026	0.951	0.959	0.955	0.966	96.85%
		DT	0	0.956	0.083	0.956	0.956	0.956	0.946	
			1	0.917	0.044	0.917	0.917	0.917	0.946	94.28%
	Bupa	L R	0	0.559	0.36	0.529	0.559	0.544	0.653	
			1	0.64	0.441	0.667	0.64	0.653	0.652	60.58%
		SVM	0	0.538	0.33	0.542	0.538	0.54	0.604	
			1	0.67	0.462	0.667	0.67	0.668	0.604	61.49%
		DT	0	0.497	0.375	0.49	0.497	0.493	0.586	
			1	0.625	0.503	0.631	0.625	0.628	0.586	57.10%

TABLE 16. Experiment with attribute selection using IG for LR, SVM and DT for three data set.

	Data Set	Algorithm	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	% Accuracy
With Attribute Selected by Information Gain	Pima	L R	0	0.88	0.429	0.793	0.88	0.834	0.832	
			1	0.571	0.12	0.718	0.571	0.636	0.832	77.22%
		SVM	0	0.898	0.459	0.785	0.898	0.838	0.72	
			1	0.541	0.102	0.74	0.541	0.625	0.72	77.34%
		DT	0	0.814	0.403	0.79	0.814	0.802	0.751	
			1	0.597	0.186	0.632	0.597	0.614	0.751	73.83%
	Wisconsin	L R	0	0.965	0.12	0.938	0.965	0.952	0.926	
			1	0.88	0.035	0.93	0.88	0.904	0.95	93.56%
		SVM	0	0.963	0.071	0.963	0.963	0.963	0.935	
			1	0.929	0.037	0.929	0.929	0.929	0.946	95.14%
		DT	0	0.952	0.145	0.926	0.952	0.939	0.962	
			1	0.855	0.048	0.904	0.855	0.878	0.962	91.85%
	Bupa	L R	0	0.538	0.325	0.545	0.538	0.542	0.593	
			1	0.675	0.462	0.668	0.675	0.672	0.594	61.74%
		SVM	0	0.524	0.305	0.555	0.524	0.539	0.61	
			1	0.695	0.476	0.668	0.695	0.681	0.61	62.32%
		DT	0	0.51	0.335	0.525	0.51	0.517	0.582	
			1	0.665	0.49	0.652	0.665	0.658	0.582	60.00%

from 0.95 to 0.963, finally the DT accuracy 91.56% to 94.56% while the $TPRate_{minority}$ increase from 0.846 to 0.925. In all experiment using the Wisconsin the variance ranking attributes selection have achieved higher accuracy in both general accuracy and targeting the minority classes.

4) VALIDATION EXPERIMENT OF BINARY CLASSIFICATION USING BUPA LIVER DISEASE DATA

Tables (13, 14, 15 and 16) contains the results of experiment for the following algorithms LR,SVM and DT using the Bupa liver disease data sets. The results in Tables 13 is using all the attributes in the data sets while Tables 14 is using the selected attributes by the peak accuracy threshold as explained in section IV-D.1 and Figure 10 is a representation of the graph of the technique to identify the peak threshold attributes.

The general accuracy of LR for increased from 54.20% to 68.12%, while the increased of the $TPRate_{minority}$ is 0.565 to 0.79. The SVM accuracy increased from 58.84% to 70.72% while the $TPRate_{minority}$ increases from 0.645 to 0.89. Finally, the DT accuracy is from 57.97% to 68.70%, the $TPRate_{minority}$ for DT increased from 0.64 to 0.80

V. DISCUSSION

There is a noticeable pattern in all the experiments which are basically an increased in the $TPRate_{minority}$ (class 1). In general there is an increase in accuracy as a result of an increase in

	Pima			
	TP(min)	Precision(min)	F-Measure(min)	% Accuracy
LR-variance Rank	0.578	0.728	0.644	77.73%
LR-PearsonC	0.571	0.718	0.636	77.21%
LR-Informatiion G	0.571	0.718	0.636	77.22%
SVM-variance Rank	0.541	0.74	0.625	77.34%
SVM-PearsonC	0.534	0.737	0.619	77.08%
SVM-Informatiion G	0.541	0.74	0.625	77.34%
DT-variance Rank	0.634	0.63	0.632	74.22%
DT-Pearson C	0.556	0.668	0.607	74.87%
DT-Informatiion G	0.597	0.632	0.614	73.83%

	Wiscosin			
	TP(min)	Precision(min)	F-Measure(min)	% Accuracy
LR-variance Rank	0.954	0.954	0.954	96.85%
LR-PearsonC	0.95	0.95	0.95	96.57%
LR-Informatiion G	0.88	0.93	0.904	93.56%
SVM-variance Rank	0.963	0.951	0.957	97.00%
SVM-PearsonC	0.959	0.951	0.955	96.85%
SVM-Informatiion G	0.929	0.929	0.929	95.14%
DT-variance Rank	0.925	0.918	0.921	94.56%
DT-Pearson C	0.917	0.917	0.917	94.28%
DT-Informatiion G	0.855	0.904	0.878	91.85%

	Bupa			
	TP(min)	Precision(min)	F-Measure(min)	% Accuracy
LR-variance Rank	0.79	0.699	0.742	68.12%
LR-PearsonC	0.64	0.667	0.653	60.58%
LR-Informatiion G	0.675	0.668	0.672	61.74%
SVM-variance Rank	0.89	0.693	0.779	70.72%
SVM-PearsonC	0.67	0.667	0.668	61.49%
SVM-Informatiion G	0.695	0.668	0.681	62.32%
DT-variance Rank	0.8	0.702	0.748	68.70%
DT-Pearson C	0.625	0.631	0.628	57.10%
DT-Informatiion G	0.665	0.652	0.658	60.00%

FIGURE 11. Summary table of comparison.

the prediction of class 1 ($TPRate_{minority}$). The emphasis here is not on the machine learning algorithm that was used but on the attribute that was selected from the variance ranking. This is to demonstrate that our technique is independent on any algorithm but dependent on the intrinsic properties of the data set particularly the measurement of central tendencies, ie the variance.

Although three algorithm LR, SVM and DT were used on three dataset; the Pima Indians Diabetes Data, Wisconsin Breast Cancer and Bupa liver diseases data, in all over twenty experiment were conducted. it shows that more accurate predictions were obtained and more minority target classes were accurately classified using the identified attributes. Additionally, there was an increase in the overall precision, recall, and F-measure. The problem associated with imbalanced data is ubiquitous and will continue to elicit customized solutions. We demonstrated a new attribute selection technique. The variance ranking attributes selection technique is significantly similar in performance to other attributes selections techniques, notably Pearson correlation and information gain which are categorized as filter attribute selection technique. We made a case why our method should be compared to the same family of filtered selection techniques by providing numerous cogent reasons such as; our method is not algorithm suggestive and does not eliminate any attribute but rather ranked them as done by other filter methods. To validate our work, three pronged approach was used.

Firstly, we performed attributes selections using our methods (variance ranking) and compare the result with that of PC and IG which are the state-of-art filter attributes selection methods, the results are in Tables 7, 8, 9 and 10. The conclusion here is that the choice of attributes selection technique depends on the context of the application and the domain of usage and that no two feature/attributes selection method thus give the same answers 100%. Hence, we encountered a new problem here, this led to the second validation technique. How could we compare or rather quantitatively grade our result with that of state-of-art filter methods?'', the PC and IG. Should we say variance ranking, PC and IG are 70%, 80% or 90% similar?. How do we calculate the similarity index?. Thus we invented a novel approach of calculating the similarity index, since the existing method like cosine, Jaccard etc similarity measurement appeared to be inadequate please see section IV-C for Ranked Order Similarity-ROS and the reasons behind it, we used it to compare our method with PC and IG. Finally, we also used the identified attributes with three binary classification models (logistic regression, support vector machine and decision tree) to filter out the most significant attributes from the datasets used (Pima Indians Diabetes and Wisconsin Breast Cancer data, Bupa liver disease data set). The results obtained are shown in Tables (13, 14, 15 and 16) an increase in capturing the minority positive classes (rare) with the selected attributes compared to the result without the selected attributes. Though in some of the results the increments are significant, while in others the increment is small, it is noteworthy that in all cases there was an increment.

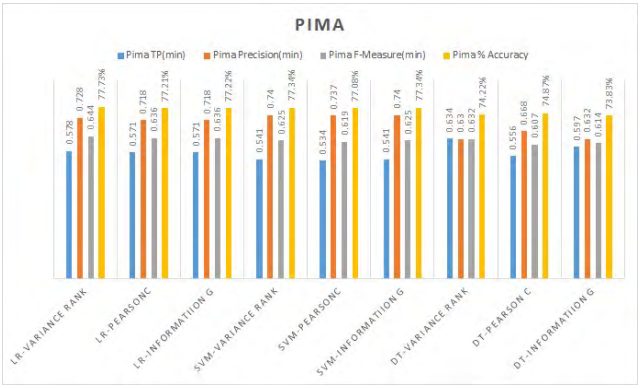

FIGURE 12. Comparison of attributes selection results for LR,SVM and DT using Pima data.

The summary table in Figure 11 and the associated graphs in figures (12,13 and 14) is a point of focus to explicitly establish the superiority of our technique (variance ranking) over PC and IG. In all the comparison eg (LR-variance Rank, LR-Pearson C and LR-Informatiion G) the variance ranking have performed better. We attributed this superior performance not only to the attributes but also to the ranking because the most significant attributes were identified and ranked earlier hence it took only 4 attributes to attained the peak accuracy threshold as against others that took more than 6 attributes.

One of the strongest benefit of attributes selection in data mining and machine learning is not only high accuracy/predictions, but achieving the predictions with

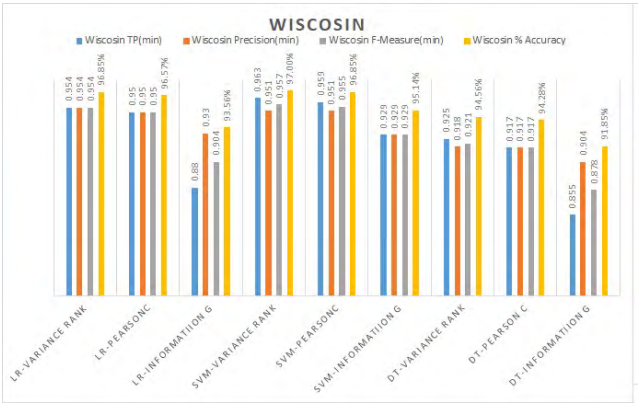

FIGURE 13. Comparison of attributes selection results for LR,SVM and DT using Wisconsin data.

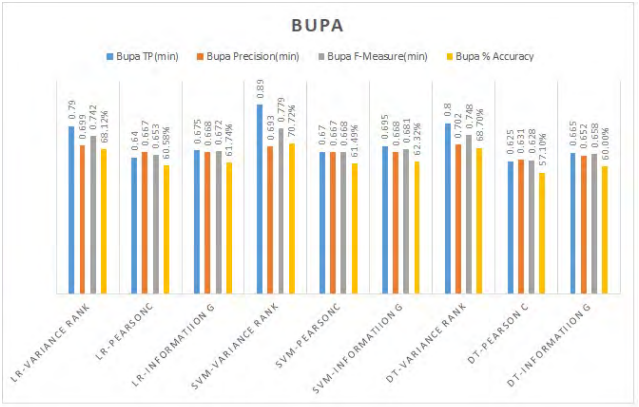

FIGURE 14. Comparison of attributes selection results for LR,SVM and DT using Bupa data.

fewer attributes. If a prediction of accuracy 80% is achieved using 10 attributes and another prediction of the same 80% is achieved using 5 attributes out of the 10 attributes the second predictions is more superior to the first in terms of resources needed to get the same predictive accuracy, apart from the accuracy another point where our techniques surpasses the others is using fewer attributes.

A. CONCLUSION AND FUTURE WORK

We reviewed the problems of imbalanced class distribution in a dataset as it relates to intrinsic characteristics. In this case, the variance of the data item and the issue of imbalanced datasets and classification algorithm, will continue to attract interest within the data science communities both in industry and academia. In our method for the proof of concept (POC), several experiments were carried out. The accuracy of the results surpassed the benchmark and was similar on some occasions. The major finding of this work can be summarized as follows:

- Variance ranking attributes selection techniques based on the intrinsic properties of each attribute in a binary classification context.

  It has long been suspected that intrinsic properties of attributes and the measure of central tendency can objectively correlate with the logical distance of the attributes to the target class. This study is the pioneer in this regard as one of the first of its kind to consider this area of attribute knowledge. This viewpoint has not been explored properly by researchers. Specifically, it has not been explored to the extent that definite conclusions could be made regarding the extent to which the intrinsic properties of attributes correlate to the target class. We have concluded that in some particular data types similar to the ones used in this research, variance does correlate to the target class in a binary context.
- A novel similarity measurement (ranked order similarity-ROS) has been invented; this technique is intended to measure the similarity between two or more sets that contain the same elements ranked in a different order. The ROS techniques are a means of grading and measuring similarities where other similarity measurement techniques are inadequate or not applicable.
- When attributes are ranked, the significant attributes are those at the peak threshold performance.

  As a subset of the data has been used, this aligned very well with the slight modification to the approach to variance when the whole population (N) and when the sample $(N - 1)$ of the whole population is being considered. In this regard, there are no conflicts; therefore, it follows that variance ranking techniques can also be applied to the whole population as well as a sample of the whole population.
- We also proposed that similarity index is an efficacious way of validating the results of experimental findings and in many instances it could be better and more dependable means of demonstrating the proof of concept (POC). The best practices in the validation of processes and experimental results should not be traditional but rigorous, invective and context based and each method of validation should be independent of each other, for example we used comparison with the bench mark, similarity index and predictive modelling and each of them are independent of the other. On each technique that was applied the evidential results point in same direction, Therefore, double-blind independent evaluation and validation (DBI-E V) techniques are highly recommended for any POC.

Similar to many attribute selection techniques, there is no single technique that can be used for all types of datasets. Depending on some intrinsic properties of the data items, each known technique must be used on the correct dataset. For example, the Pearson correlation technique cannot be used for categorical datasets.

Hence, this variance ranking attributes selection has the following limitations.

- The variable must be numeric (discrete or continuous).
- The variable must not be categorical.

Our next work hopes to solve the limitations as itemized above and to apply this technique to a multiclass dataset. We hope to ascertain the extent and threshold to which the intrinsic properties of the data item affect the attributes and generally, the learning algorithm. Additionally, we will explore the relationships of the measurements of central tendency to the significant attributes, if any.

February 05, 2019

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

REFERENCES

[1] R. Longadge and S. Dongre, "Class imbalance problem in data mining: Review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, p. 1707, 2013.
[2] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.
[3] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.
[4] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
[5] B. A. G. Nguyen Hoang and S. Phung, "Learning pattern classification tasks with imbalanced data sets," in *Pattern Recognition*, P.-Y. Yin, Ed., Vukovar, Croatia: InTech, 2009.
[6] A. Maratea, A. Petrosino, and M. Manzo, "Adjusted F-measure and kernel scaling for imbalanced data learning," *Inf. Sci.*, vol. 257, pp. 331–341, Feb. 2014.
[7] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[8] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.
[9] C. Bessiere, K. Stergiou, and T. Walsh, "Domain filtering consistencies for non-binary constraints," *Artif. Intell.*, vol. 172, nos. 6–7, pp. 800–822, 2008.
[10] E. Fagiuoli and M. Zaffalon, "2U: An exact interval propagation algorithm for polytrees with binary variables," *Artif. Intell.*, vol. 106, no. 1, pp. 77–107, Nov. 1998.
[11] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Making*, vol. 11, no. 1, p. 51, 2011.
[12] A. S. Hussein, W. M. Omar, X. Li, and M. Ati, "Efficient chronic disease diagnosis prediction and recommendation system," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2012, pp. 209–214.
[13] S. M. Oh, K. M. Stefani, and H. C. Kim, "Development and application of chronic disease risk prediction models," *Yonsei Med. J.*, vol. 55, no. 4, pp. 853–860, 2014.
[14] W. Lee, C.-H. Jun, and J.-S. Lee, "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification," *Inf. Sci.*, vol. 381, pp. 92–103, Mar. 2017.
[15] Analytics Vidhya Content Team. (Mar. 2016). *Practical Guide to Deal With Imbalanced Classification Problems*. Accessed: Mar. 19, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems
[16] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[18] J. Brownlee. (Aug. 18, 2015). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. Accessed: Apr. 28, 2018. [Online]. Available: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
[19] G. M. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?" in *Proc. DMIN*, vol. 7, Jun. 2007, pp. 35–41.
[20] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Proc. Workshop Learn. Imbalanced Datasets II*, Washington, DC, USA, vol. 11, 2003, pp. 1–8.
[21] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Improving software-quality predictions with data sampling and boosting," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 39, no. 6, pp. 1283–1294, Nov. 2009.
[22] A. Liu, J. Ghosh, and C. E. Martin, "Generative oversampling for mining imbalanced datasets," in *Proc. DMIN*, 2007, pp. 66–72.
[23] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145–9154, 2016.
[24] Z. Yang and D. Gao, "An active under-sampling approach for imbalanced data classification," in *Proc. 5th Int. Symp. Comput. Intell. Design (ISCID)*, vol. 2, Oct. 2012, pp. 270–273.
[25] A. Estabrooks, T. H. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, 2004.
[26] M. Altini. (Aug. 17, 2015). *Dealing With Imbalanced Data: Undersampling, Oversampling and Proper Cross-Validation*. [Online]. Available: https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation
[27] P. Li, T.-G. Liang, and K.-H. Zhang, "Imbalanced data set CSVM classification method based on cluster boundary sampling," *Math. Problems Eng.*, vol. 2016, Jun. 2016, Art. no. 1540628. [Online]. Available: https://www.hindawi.com/journals/mpe/2016/1540628/abs/
[28] J. Song, X. Huang, S. Qin, and Q. Song, "A bi-directional sampling based on K-means method for imbalance text classification," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–5.
[29] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, 2016.
[30] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Comparison of data sampling approaches for imbalanced bioinformatics data," in *Proc. FLAIRS Conf.*, 2014, pp. 1–4.
[31] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 4, pp. 42–47, 2012.
[32] R. Batuwita and V. Palade, "Class imbalance learning methods for support vector machines," Singapore-MIT Alliance Res. Technol. Centre, Univ. Oxford, Oxford, U.K., Tech. Rep., 2013.
[33] M. Graff and R. Poli, "Practical performance models of algorithms in evolutionary program induction and other domains," *Artif. Intell.*, vol. 174, no. 15, pp. 1254–1276, 2010.
[34] A. Kibekbaev and E. Duman, "Benchmarking regression algorithms for income prediction modeling," *Inf. Syst.*, vol. 61, pp. 40–52, Oct. 2016.
[35] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Proc. ICML Workshop Learn. Imbalanced Data Sets II*, Washington, DC, USA, 2003, pp. 49–56.
[36] G. Ditzler and R. Polikar, "Incremental learning of concept drift from streaming imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2283–2301, Oct. 2013.
[37] H. Dubey and V. Pudi, "Class based weighted k-nearest neighbor over imbalance dataset," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2013, pp. 305–316.
[38] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Syst. Appl.*, vol. 91, pp. 464–471, Jan. 2018.
[39] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.
[40] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, vol. 17, no. 1, pp. 973–978.
[41] M. J. Siers and M. Z. Islam, "Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem," *Inf. Syst.*, vol. 51, pp. 62–71, Jul. 2015.
[42] T. Wang, Z. Qin, S. Zhang, and C. Zhang, "Cost-sensitive classification with inadequate labeled data," *Inf. Syst.*, vol. 37, no. 5, pp. 508–516, 2012.
[43] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2010, pp. 1–8.

[44] M. Lango and J. Stefanowski, "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data," *J. Intell. Inf. Syst.*, vol. 50, no. 1, pp. 97–127, 2017.
[45] H. Yin, K. Gai, and Z. Wang, "A classification algorithm based on ensemble feature selections for imbalanced-class dataset," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. (HPSC), IEEE Int. Conf. Intell. Data Secur. (IDS)*, Apr. 2016, pp. 245–249.
[46] P. D. Joyce. (2014). *Expectation and Variance for Continuous Random Variables*. Accessed: Mar. 13, 2018. [Online]. Available: https://mathcs.clarku.edu/~djoyce/ma217/contexp.pdf
[47] Wyzant. (May 2016). *Variance and Standard Deviation of a Random Variable*. Accessed: May 13, 2018. [Online]. Available: https://www.wyzant.com/resources/lessons/math/statistics_and_probability/expected_value/variance
[48] J. Orloff and J. Bloom. (2014). *Expectation, Variance and Standard Deviation for Continuous Random Variables*. Accessed: May 15, 2018. [Online]. Available: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6a.pdf
[49] Yale University. (2017). *Mean and Variance of Random Variables*. Accessed: Jun. 10, 2018. [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/rvmnvar.htm
[50] J. Orloff and J. Bloom. (2014). *Manipulating Continuous Random Variables*. Accessed: May 23, 2018. [Online]. Available: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading5d.pdf
[51] (Jun. 6, 2014). *Probability Distributions*. Accessed: Nov. 24, 2018. [Online]. Available: https://www.le.ac.uk/users/dsgp1/COURSES/LEISTATS/STATSLIDE4.pdf
[52] S. Erlander, "On the relationship between the discrete and continuous models for combined distribution and assignment," *Transp. Res. B, Methodol.*, vol. 22, no. 5, pp. 371–382, 1988.
[53] (2014). *Mixed Distributions*. Accessed: Sep. 5, 2018. [Online]. Available: http://math.bme.hu/~nandori/Virtual_lab/stat/dist/Mixed.pdf
[54] H. Pishro-Nik. (2016). *Introduction to Probability, Statistics and Random Processes*. Accessed: Mar. 2, 2018. [Online]. Available: http://math.bme.hu/~nandori/Virtual_lab/stat/dist/Mixed.pdf
[55] S. Bruce, Z. Li, A. H. Chieh, and S. Mukhopadhyay, "Nonparametric distributed learning architecture for big data: Algorithm and applications," *IEEE Trans. Big Data*, to be published.
[56] F. S. Nahm, "Nonparametric statistical tests for the continuous data: The basic concept and the practical use," *Korean J. Anesthesiol.*, vol. 69, no. 1, pp. 8–14, 2016.
[57] S. Guo, S. Zhong, and A. Zhang, "Privacy-preserving Kruskal–Wallis test," *Comput. Methods Programs Biomed.*, vol. 112, no. 1, pp. 135–145, 2013.
[58] S. García, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization," *J. Heuristics*, vol. 15, no. 6, p. 617, 2009.
[59] N. L. Leech and A. J. Onwuegbuzie, "A call for greater use of nonparametric statistics," Annu. Meeting Mid-South Educ. Res. Assoc., Chattanooga, TN, USA, Tech. Rep., Nov. 2002.
[60] F. Musyimi. (May 5, 2017). *Kruskal Wallis H Test: Definition, Examples & Assumptions*. Accessed: May 23, 2018. [Online]. Available: http://www.statisticshowto.com/kruskal-wallis/
[61] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.
[62] R. Lowry. (2018). *Concepts and Applications of Inferential Statistics-The Kruskal-Wallis Test*. Accessed: May 20, 2018. [Online]. Available: http://vassarstats.net/textbook/ch14a.html
[63] B. Delgutte. (2007). *Random Variables and Probability Density Functions-Biomedical Signal and Image Processing*. Accessed: May 27, 2018. [Online]. Available: http://web.mit.edu/~gari/teaching/6.555/lectures/ch_pdf_sw.pdf
[64] Wikipedia. *F-Distribution*. Accessed: Dec. 25, 2018. [Online]. Available: https://en.wikipedia.org/wiki/F-test
[65] Introduction to Statistics. *The F Distribution and the F-Ratio*. Accessed: Dec. 25, 2018. [Online]. Available: https://courses.lumenlearning.com/introstats1/chapter/the-f-distribution-and-the-f-ratio/
[66] Yourself Statistics. *F Distribution*. Accessed: Dec. 25, 2018. [Online]. Available: https://stattrek.com/probability-distributions/f-distribution.aspx
[67] J. Clark. *F-Distribution Explained*. Accessed: Apr. 28, 2018. [Online]. Available: https://magoosh.com/statistics/f-distribution-explained/
[68] D. G. Altman and J. M. Bland, "Parametric v non-parametric methods for data analysis," *Bmj*, vol. 338, p. a3167, Apr. 2009.
[69] *Properties of Variance*. Accessed: Dec. 25, 2018. [Online]. Available: https://courses.cs.washington.edu/courses/cse312/13wi/slides/var+zoo.pdf
[70] I. Ishida and R. F. Engle, "Modeling variance of variance: The square root, the affine, and the CEV GARCH models," Dept. Finances, New York, NY, USA, Working Papers, 2002.
[71] *Diabetes America*, Nat. Inst. Health, Nat. Inst. Diabetes Digestive Kidney Diseases, Bethesda, MD, USA, 1995.
[72] R.-E. Fan and C.-J. Lin. (May 5, 2017). *LIBSVM Data: Classification, Regression, and Multi-Label*. Accessed: Jun. 10, 2018. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
[73] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide, USA: SPSS Inc," *CRISPWP-0800*, 2000.
[74] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.
[75] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proc. IEEE Int. Conf. Privacy, Secur. Data Mining*, vol. 14, 2002, pp. 1–8.
[76] U. M. Fayyad and K. B. Irani, "The attribute selection problem in decision tree generation," in *Proc. AAAI*, 1992, pp. 104–110.
[77] H. Schmid, "Probabilistic part-ofispeech tagging using decision trees," in *New Methods Language Processing*. London, U.K.: Routledge, 2013, p. 154.
[78] F. E. Harrell, "Ordinal logistic regression," in *Regression Modeling Strategies*. Springer, 2015, pp. 311–325.
[79] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Application of Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.
[80] A. J. Barros and V. N. Hirakata, "Alternatives for logistic regression in cross-sectional studies: An empirical comparison of models that directly estimate the prevalence ratio," *BMC Med. Res. Methodol.*, vol. 3, no. 1, p. 21, 2003.
[81] (May 2018). *The Logistic Regression Algorithm*. Accessed: Nov. 18, 2018. [Online]. Available: https://machinelearning-blog.com/2018/04/23/logistic-regression-101/
[82] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.
[83] C. Qi, A. Fourie, G. Ma, X. Tang, and X. Du, "Comparative study of hybrid artificial intelligence approaches for predicting hangingwall stability," *J. Comput. Civil Eng.*, vol. 32, no. 2, 2017, Art. no. 04017086.
[84] C. Qi, A. Fourie, X. Du, and X. Tang, "Prediction of open stope hangingwall stability using random forests," *Natural Hazards*, vol. 92, no. 2, pp. 1179–1197, 2018.
[85] V. López, A. Fernández, and F. Herrera, "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed," *Inf. Sci.*, vol. 257, pp. 1–13, Feb. 2014.
[86] C. Qi, A. Fourie, G. Ma, and X. Tang, "A hybrid method for improved stability prediction in construction projects: A case study of stope hangingwall stability," *Appl. Soft Comput.*, vol. 71, pp. 649–658, Oct. 2018.
[87] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[88] S. Sasikala, S. A. A. Balamurugan, and S. Geetha, "Multi filtration feature selection (MFFS) to improve discriminatory ability in clinical data set," *Appl. Comput. Inform.*, vol. 12, no. 2, pp. 117–127, 2016.
[89] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[90] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Tech. Rep., 2000.
[91] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, 2013.
[92] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, pp. 59–81, May 2014.
[93] N. Hoque, D. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6371–6385, 2014.

[94] W. Dai and W. Ji, "A mapreduce implementation of C4.5 decision tree algorithm," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 49–60, 2014.
[95] R. T. Ng and J. Han, "E cient and E ective clustering methods for spatial data mining," in *Proc. VLDB*, 1994, pp. 144–155.
[96] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[97] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*. Springer, 2006, pp. 25–71.
[98] S. Polamuri. (Jun. 6, 2015). *Implementing the Five Most Popular Similarity Measures in Python*. Accessed: Sep. 21, 2018. [Online]. Available: http://bigdata-madesimple.com/implementing-the-five-most-popular-similarity-measures-in-python/
[99] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.
[100] M. M. Rahman, "Machine learning based data pre-processing for the purpose of medical data mining and decision support," Ph.D. dissertation, Univ. Hull, Hull, U.K., 2014.
[101] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, Nov./Dec. 2003.
[102] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manage.*, vol. 2, no. 2, pp. 271–277, 2010.

SOLOMON H. EBENUWA received the B.Ed. degree in physics/science and the M.Sc. degree in advanced computing (data mining). He is currently pursuing the Ph.D. degree with the School of Architecture, Computing and Engineering, University of East London. He was a Physics and Mathematics Teacher in Sixth form colleges and further education institutions within and outside London. He is currently a Lecturer with the Newham College of Further Education, London. His research interests include data mining, machine learning, decisions based systems, and programming data structure in relation to machine learning algorithms. He is a Fellow of the U.K. Higher Education Academy.

SAEED SHARIF received the Ph.D. degree from Brunel University London. He is currently an Associate Professor with the School of Architecture, Computing and Engineering. He is working closely with clinicians and policy makers nationally and internationally to improve the clinical settings and the healthcare systems. His research interests include innovative tele-health, medical technology, digital health care and medical assistive technology, artificial intelligence, medical image analysis and visualization, intelligent diagnosis systems, smart biomedical image, and bio signal acquisition (MRI, PET, and EEG).

He is a Fellow of the U.K. Higher Education Academy. He received many academic and research awards. He has participated in many national and international conferences, e.g., ICIP. He has served as a Reviewer for many journals, e.g., IET IPJ and Elsevier CBMJ.

MAMOUN ALAZAB received the Ph.D. degree in computer science from the School of Science, Information Technology and Engineering, Federation University of Australia. He is currently an Associate Professor with the College of Engineering, IT and Environment, Charles Darwin University. He is also a Cyber Security Researcher and a Practitioner with industry and academic experience. He organized and participated more than 70 conferences and workshops, seven as a General Chair. His research interests include cyber security, blockchain technologies, and digital forensics of computer systems, particularly cybercrime detection and prevention. He has more than 100 research papers. He presented many invited and Keynotes talks at conferences and venues (22 events in 2018 alone). He is an Editor on multiple editorial boards, including an Associate Editor of the IEEE Access, an Editor of the *Security and Communication Networks Journal* and a Book Review Section Editor: the *Journal of Digital Forensics, Security and Law*. He has been involved in past research work to support agencies like the Australian Federal Police, Attorney General's Department, and major banks in Australia.

AMEER AL-NEMRAT is currently an Associate Professor with the School of Architecture, Computing and Engineering, University of East London (UEL). He is also the Director for the Professional Doctorate in information security and the M.Sc. information security and computer forensics programmes. In addition, he is also the Founder and the Director of the Electronic Evidence Laboratory, UEL, where he closely working with Law and Policy maker, National and International, agencies on cyber security research projects. His research interest includes cyber security and digital forensics has had considerable impact on public and industry practice at national and international level. His research has influenced range of organizations and agencies, including the U.K. Government and EU Council. He personally involved in the professionalization of Initial police training in Africa, South America, and Middle East. His experience and opinion were directly required to advise and help to maintain International cyber security.