JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SC: A Sampling from Clusters for Reduction of Dataset Size

[1]Onima Tigga, [2]Jaya Pal, [3] Debjani Mustafi
[123]Assistant Professor
[1]Department of Computer Science & Engg.
[1]Birla Institute of Technology, Ranchi, India

Abstract: Since managing enormous datasets in the real world is difficult, it is necessary to minimize the size of the data set, so that the accuracy of the original dataset is no longer impacted. In this study, the categorization of the white wine dataset is examined using a number of machine learning techniques, including Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), and Logistic Regression (LR). Additionally, we utilized the stated dataset using the defined methodologies and presented the Sampling from Clusters (SC) approach. The white wine dataset is first clustered using our suggested method SC, and then 95% of the data from each cluster is removed and combined to create a standard dataset for classification process. For 90%, 85%, and 80% of the data, the same procedure is repeated. On the other hand, we used a random sampling (RS) technique to work with 95% of the data from the dataset in question, and we compared the results with SC using evaluation metrics like accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC), Area under the Curve (AUC), binomial confidence interval (CI), and MSE. With 90%, 85%, and 80% of the datasets, the same procedure is repeated. According to statistics, confidence intervals CI become tighter as the quantity of test data N increases; they range from 0.72 to 0.76 for NB, 0.73 to 0.79 for SVM, 0.82 to 0.86 for RF, 0.75 to 0.77 for KNN, and 0.74 to 0.80 for LR.

Keywords- *KNN, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine.*

I. INTRODUCTION

Data mining is a method for extracting important information from vast amounts of data. The data is then translated into this information using a variety of techniques. It ought to be suitable for all types of data repositories [1-2]. So that everyone may focus on cleaning the data for feasibility, data should be cleansed before further processing [3]. The most common application of machine learning is classification approach. Machine learning techniques encompass Naïve Bayes, K nearest Neighbour, Support Vector Machine, Decision Tree, and more. It maps data into predefined classes [2]. Earlier researchers investigated the performances of various machine learning techniques rather than optimization, these studies concentrated on their specific impact. Some researchers tried out hybrid optimization methods [4]. In modern era, the consumption of wine correlates with rate variability. With increased consumption, the wine industry is allowing to provide high-quality wine at a lower cost. Although the majority of the chemicals in various types of wines are the same. Nowadays, classifying different wines is critical for quality assurance [5]. Wine quality is one of the precious factors of wine industries. The fragrance and flavor characteristics of the wine are important factors in determining its quality. [6-7]. Due to security and strategic concerns, only physiochemical (input) and output features are available [8-9]. Wine quality is determined not only by the amount of alcohol in it but also by other factors. These qualities change over time, and thus the quality of wine is refined. A physiochemical test is a laboratory test that does not require any human expertise. Several machine learning techniques are applied to predict wine quality [10]. The graphical abstract of our research work is depicted in Fig 1. For classification using labeled data; we have used five machine learning techniques: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), and Logistic Regression (LR).

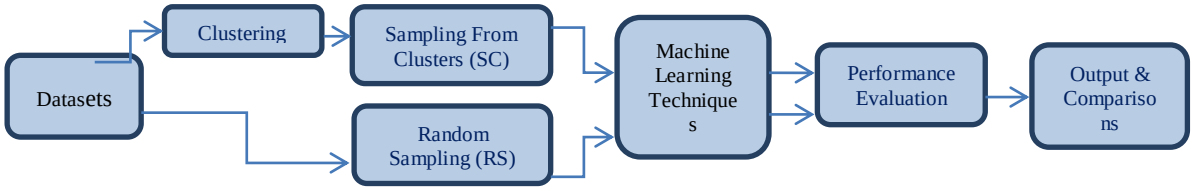

Fig. 1 Graphical Abstract of Our Research Work

The following are the primary contributions of this research work:

- Propose a new approach based on clustering, SC (Sampling from Clusters), apply all the mentioned techniques to data clusters obtained from the K-Means clustering algorithm with (95%, 90%, 85%, and 80%) of data.
- Calculate the accuracy, precision, recall, F1-score, ROC, AUC, MSE, and CI of each specified machine learning model.

- NB, SVM, RF, KNN, and LR are used with randomly selected data (95%, 90%, 85%, and 80%), and machine learning performance metrics like accuracy, precision, recall, F1-score, ROC, AUC, MSE, and CI are also calculated.
- Compare the newly proposed approach SC with the random Sampling (RS) methods to determine which produces the best outcomes.

This study aims to evaluate five machine learning classifiers to see if the classification result can be enhanced by expanding the dataset. Precision, recall, accuracy, F1-score, ROC, AUC, MSE, and CI have all been used to evaluate each classifier's performance and perform preprocessing such as correlation for feature selection, binarization for selecting two outcomes and data standardization.

In this research, each feature has been altered to have the characteristics of a normal distribution with a mean μ and standard deviation σ of 0 and 1, respectively.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

The remaining section of the paper is arranged as follows: Section 2 deals with related work, Section 3 with classification methods, and Section 4 with datasets, dataset implementations, and evaluation parameters. Our findings demonstrate that smaller data sets with extracted features perform tasks more effectively. Section 5 contains the results and analysis. Section 6 shows the improvement Analysis. Section 7 of the study wraps up with a discussion on future research.

II. RELATED WORK

Earlier, much work was done by various researchers in the area of machine learning techniques for evaluating performance. They achieved various levels of accuracy by employing various machine-learning techniques. Some of them are as follows:

In [6] V. A. Parvathy et al explained the different machine learning techniques used to evaluate how well they performed on the basis of accuracy, precision, recall, F1-score, ROC, and RF and found that they performed well. They employed datasets for red wine and white wine, machine learning techniques, and physio-chemical and chemical elements for predicting wine quality. O. D. Akanbi et al. in [7] used machine learning techniques to build models and achieved RF, as a better model when cross-validated in 10-fold. Alcohol is the most important feature variable in wine quality, while chloride and volatile acidity are the least important. In [8] Y. Er et al. used principal component analysis in feature selection to improve classification success rates in RF algorithms. In [9] Kumar S. et al. have found the increasing demand of customer's value wine, so the wine sector is investigating new advancements in both wine production and offering structures. N. Korade in [10] applied different machine learning techniques and compared them to identify the best suitable for the prediction of wine by selecting features. In [4] S. U. Ahsaan et al proposed in a study that concentrated on the variability issue associated with big data, a hybrid Support Vector Machine classifier was built. It relied on simple SVM, inconsistent Euclidean overlap metric, and in order to group and categorize data based on numerical and nominal numbers, Euclidean distance has been used. Based on the examined dataset for Amazon Unlocked mobile phones, M. Guia et al. in [12] offered a comparison of four techniques: NB, SVM, DT, and RF. The SVM had an outstanding F1 score, accuracy, precision, and recall, in accordance with the consequences. It also assessed how the brand price of mobile phones affected the polarity review. According to A. Mishra et al. in [13] on unlabelled data, self-supervised pre-Training focuses on how features are represented in learning, including contrastive loss. It was found that the suggested approach outperformed random or Image Net initialization. As basis models, they used three ResNet models of varying sizes, and experiments were done to see what would happen if the augmentation techniques were changed to create samples that are positive and negative for self-supervised training . A.H. Wibowo et al in [15] examined three different techniques —KNN, NB, and DT—that are used to reliably predict crimes and criminal activity in the Sleman Regency. The outcomes of this study indicated that the NB look had the highest accuracy.

According to T. R. Patil et al. in [16], when classifying data and evaluating effectiveness, classifier methods built on the TP rate, FP rate, and J48, which rely on decision Trees, outperform Naive Bayes. M. C. Untrono et al. in [17] investigated DT, KNN, NB, and SVM using MWMOTE (The Majority Weighted Minority Oversampling Technique). These suggestions outcomes enabled us to synthesize synthetic data more accurately while lowering the level of bias or noise. The maximum accuracy was attained by DT. S. Karthika et al. in [18] proposed using the benchmark Naive Bayesian classification algorithm to classify educational qualifications. The proposed approach performed better in accuracy tests when compared to SVM, RF, and KNN. The Naive Bayes Classifier was discussed by N. Sharma in [19] as part of the classification procedure. This probabilistic classifier is the simplest one available for classifying texts. In [20] S. Uddin et al. predicted a single disease using multiple supervised machine-learning techniques. The SVM algorithm was found to be the most commonly used algorithm (in 29 studies), after the NB techniques (in 23 studies). In contrast, the RF demonstrated greater precision. V. Sheth et al. in [21] used four classification models—DT, SVM, NB, and KNN—and five different datasets in their study. Among different algorithms, the NB is the most successful. P. T. Noi et al. in [22] used sentinel-2 image data to investigate and compare the efficiency of RF, KNN, and SVM for categorizing land use and coverage. All three classifiers displayed a comparable and good overall accuracy where the training sample size was sizable. Chen et al. in [11] devised a technique for estimating wine grade that depends on human flavour evaluations. They employed a hierarchical clustering method and an association rule method to analyze the assessments and forecast the wine grade, and found the fact that their prediction was 85.25% accurate. S.M.H.S. Iqbal in [23] utilized a technique called univariate feature selection to select the most important features from the dataset's features. P. Bhardwaj et al. in [24] predicted wine quality. To address this issue, they developed synthetic data with properties similar to those of the actual data and applied a variety of feature selection techniques. In [25] M.A. Mabayoje et al. compared three classification algorithms: DT (J-48), RF, and NB. They applied a ten-fold cross-validation technique in mandate to ponder the classifiers' enactment.

III. CLASSIFICATION METHODS

The Proposed work is focused on machine learning techniques NB, SVM, RF, KNN, and LR. The functionality of these techniques is summarized in the following section.

3.1 Naïve Bayes

For classification tasks involving continuous feature values, the classifier Naive Bayes is employed. The naïve technique gains its name from the notion that the availability of any attribute is completely self-regulating of its existence. Bayes theorem, which asserts that the likelihood of hypothesis A occurring when data B is already known is the same as the likelihood of hypothesis A occurring when data B is unknown [14-16]. A feature vector comprising n dimensions, $X=(x_1, x_2,..., x_n)$, is used to represent each sample of data in this classifier. Consider the m classes $C_1, C_2,..., C_m$. The classifier selects an unidentified data instance, X, from the class bearing the largest posterior probability. The NB classifier allocates a previously unknown sample X to class C_l if and only if $P(C_l / X) > P(C_k / X)$ for $1 \leq k \leq m$ [25].

3.2 Support Vector Machine

Using supervised machine learning, the SVM classifier can categorize data. It displays each data point in n-dimensional space with 'n' features, and the corresponding coordinate value associated with every data point is the numerical value of the feature. Finally, classification is achieved by locating hyper planes that differentiate the various classes; if a test data point can be located within a certain hyper plane, it will share the same class as the data points in its proximity [26-29].

3.3 Random Forest

The RF ensemble training technique is used to create a classifier model that divides the dataset into a number of smaller-scale datasets [30-31]. In the RF, Bootstrap aggregation is used for it is proved to be better in most situations of reducing variance to reduce the probability of over-fitting. Based on these datasets, the ensemble models are built. Then, RF generates several decision trees without pruning, forming a forest that owns a predicted value for a particular data sample. The value predicted to be the majority voting of trees. RF is a model based on a decision tree model [32]. Random forest models, which are based on the concept of decision trees, have produced an 'n' number of trees that, when compared to a single tree, have significantly increased prediction accuracy; without replacement, a random sample of every tree in the training set is taken. Decision trees tend to be tree-like structures with a root or deciding node at the top [33-34].

3.4 K Nearest Neighbour

The KNN classifier was created on learning by training tuples. A point in n-dimensional space is represented by each tuple. An n-dimensional pattern space contains all training tuples [35]. Using the given dataset, the learning and prediction analysis is performed well. In KNN predictive data mining model bases its predictions solely on closest neighbour data values without any assumption on the dataset. In this model, the 'K' is the number of nearest neighbour data values. The KNN technique decides how to classify the provided dataset based on this "K" value, or the quantity of nearest neighbours [35-36]. The training dataset is categorized directly by the KNN model. It means that the classification of a new instance is based on the class of highest tuples and that the prediction of a new instance is made by seeking the like 'K' neighbour instances in the whole training set. A similar tuple is determined using the Euclidean distance formula [37-38].

The Euclidean distance d between two points x and y in two or higher dimensions can be determined by the following formula:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (2)$$

Where n is the number of dimensions, and x_k and y_k are the kth features of x and y.

3.5 Logistic Regression

Logistic Regression is a technique of classification that establishes a decision border or the hyper plane among all the classifications in the data. This result is predicted using a linear equation incorporating with coefficients and values of input. Maximum likelihood estimation is employed to compute the coefficients. It is based on logistic or sigmoid functions [30, 32-33]. Any real-valued number can be converted to its inverse $\{-\infty, \infty\}$, to the range $\{0, 1\}$ using the sigmoid function. In binary classification, the LR model assesses the likelihood of the majority class. For instance, if LR predicts a probability of 0.1 for the default class and predicts a probability of 0.9 for the second class, the query point might be categorized as being in the second class [39].

IV. EXPERIMENTS

4.1 Datasets

This study made use of a dataset from the UCI Machine learning repository [40]. However, as shown in Table 1, we created our datasets by using the White Wine dataset and the K-Means clustering method. The original dataset of White wine used in this study has 4898 data instances and was randomly taken (95%, 90%, 85%, and 80%) data from each cluster for preprocessing.

The complete workflow diagram of research work is shown in Fig 2.

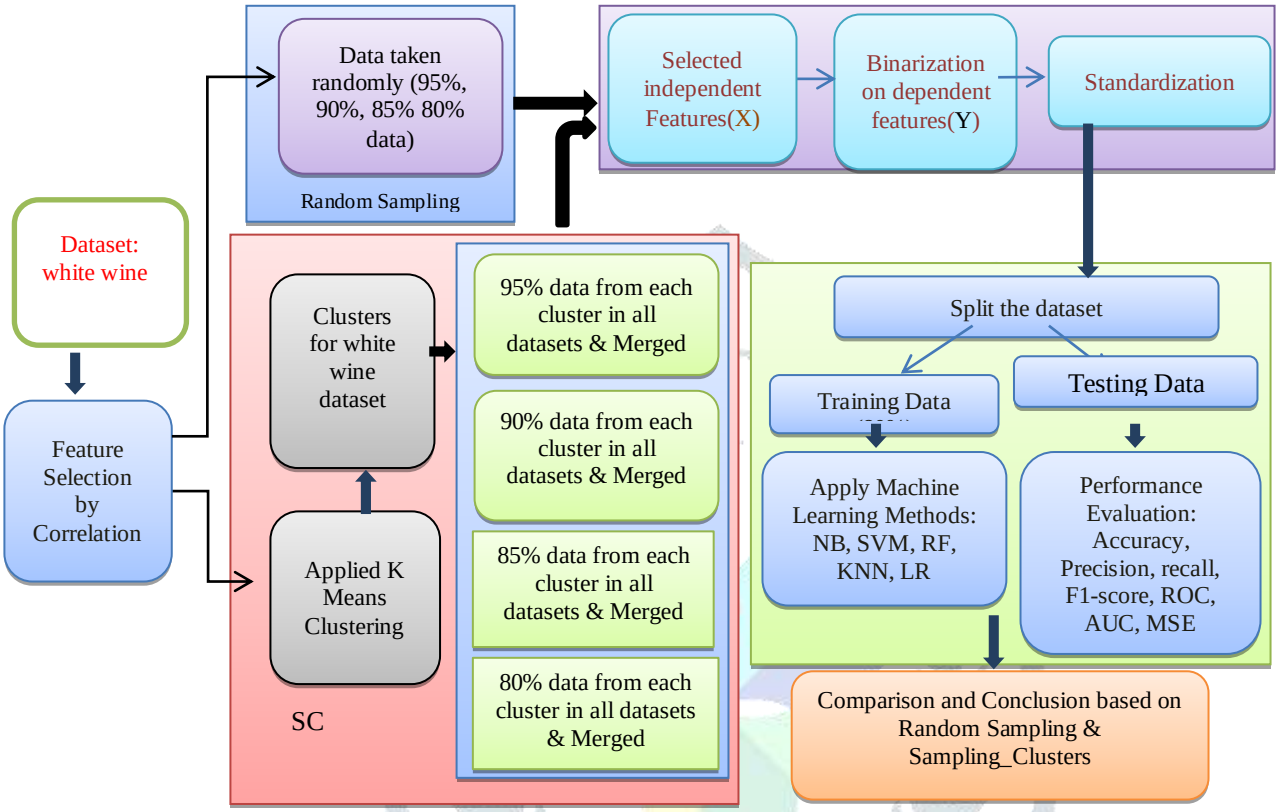

Fig. 2 Complete workflow diagram of research work

a) *Feature Selection*

Features have been extracted from the dataset White Wine (11 attributes) by taking a threshold value of 0.6 with positive correlation. On using this threshold value, seven attributes are extracted such as fixed acidity, volatile acidity, citric acid, chloride, pH, sulphates, and alcohol.

b) *Random Sampling (RS)*

The original dataset of white wine used in this study has 4898 data instances and randomly taken (95%, 90%, 85%, and 80%) of data for preprocessing.

c) *Sampling from Clusters (SC)*

In our proposed approach SC, we first created Clusters by applying K-Means clustering. By the elbow curve, minimum error was found at 8 clusters. The size of the data clusters for the dataset is shown in Table 1 and Fig 3.

Table 1 Sample Sizes of Data Subsets

Dataset	Size of the Dataset	No. of Clusters	Size of each cluster
White Wine	4898	8	C1 (602) C2 (415) C3 (1000) C4 (464) C5 (406) C6 (737) C7 (407) C8 (867)

Clusters of white wine dataset are shown in Fig 3:

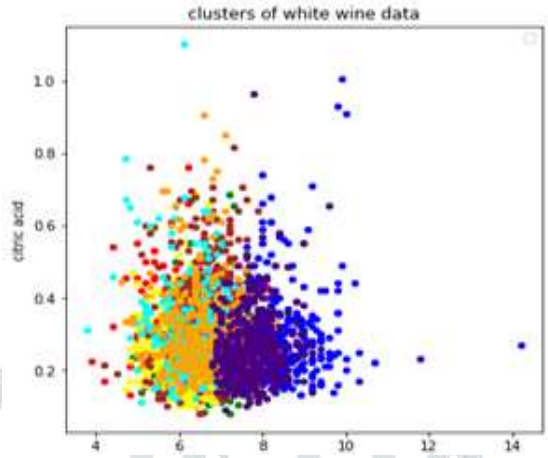

Fig. 3 Clusters of White Wine Dataset

Here, 95%, 90%, 85%, and 80% data are taken from each cluster and formed reduced data subsets for the white wine dataset. These data subsets are used for further processing. Then, selected attributes are considered as independent features (X) and class is taken as a dependent feature (Y) and applied Binarization on Y. Normalization was applied and data was split into training testing with 80:20 ratios. Machine learning techniques (NB, SVM, RF, KNN, and LR) are applied to the Training and Testing data sets.

d) *Performance Evaluation & Results*

We compare the proposed approach SC with the randomly selected RS approach. The widely used metrics are Accuracy, precision, recall, F1-score, ROC, AUC, MSE and CI. These measures with higher values indicate better performance with lower MSE value.

4.2 Evaluation Metrics

To validate how well the suggested Clustering-based approach SC performs for classification techniques and a comparison with NB, SVM, RF, KNN, and LR techniques for (95%, 90%, 85%, and 80%) of data, for the white wine dataset, we employed accuracy, precision, recall, F1-score, ROC, AUC, MSE and binomial confidence interval CI.

4.2.1 Accuracy

Accuracy is defined as a percentage of accurate predictions to the total amount of data points used in an instance.

$$Accuracy = \frac{Numberofcorrectpredictions}{Tota\ln umberofpredictionsmade} \tag{3}$$

4.2.2 Precision

The percentage of accurately predicted positive outcomes.

$$\Pr ecision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4}$$

4.2.3 Recall

The recall is calculated by dividing the total number of true positives by the total number of true positives and false negatives.

$$\Re call = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5}$$

4.2.4 F1-Score

The average of recall and precision is used to determine the F1 score.

$$F1 - score = 2 * \frac{\Pr ecision * \Re call}{\Pr ecision + \Re call} \tag{6}$$

Several machine learning techniques are applied to predict wine quality [10]. The graphical abstract of our research work is depicted in Fig 1.For classification using labeled data; we have used five machine learning techniques: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), and Logistic Regression (LR).

Several machine learning techniques are applied to predict wine quality [10]. The graphical abstract of our research work is depicted in Figure 1.For classification using labeled data; we have used five machine learning techniques: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), and Logistic Regression (LR).

4.2.5 ROC (Receiver Operating Char) Curve

The ROC curve is a graph that illustrates the true positive rate (y-axis) and the number of false positives (x-axis) at various threshold stages. It exhibits the relationship between the prevalence of false positives and true positives. A perfect predictor would have one true positive and zero false positives. The percentage of positive tuples that the classifier correctly identifies as optimistic is referred to as the true positive rate. Negative tuples with incorrect labels result in false positive rates.

4.2.6 AUC (Area Under The Curve)

AUC is referred to as the integrated area of the True Positive Rate vs. False Positive Rate graph. The area under the curve (AUC) is only one value that measures how well the model distinguishes the labels.

4.2.7 MSE

The MSE is the mean of the squared differences that exist between anticipated and actual outcomes (Pi and Ai). It is calculated by using the equation given in (7), where n is the number of tuples [42].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2 \quad (7)$$

4.2.8 Binomial Confidence Interval For Accuracy

A test set has M tuples, the correctly predicted tuples by the model is Y and the true accuracy of the model is s. If the prediction task is considered as a binomial distribution, then Y has a binomial distribution with mean (M * s) and variance (M * s (1-s)). The empirical accuracy, accuracy = Y/M, has a binomial distribution with mean s and variance (s * (1-s)/) M. When M is large, the binomial distribution is treated as a normal distribution. The confidence interval for accuracy based on normal distribution is as follows:

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{accuracy-s}{\sqrt{s(s-1)/M}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \quad (8)$$

Where $Z_{\alpha/2}$ = upper bound and $Z_{1-\alpha/2}$ = lower bound, when the confidence level is (1- α) obtained from a standard normal distribution. It shows $Z\alpha_{/2} = Z_{1-\alpha/2}$ if Z=0, since the normal distribution is symmetrical as expected. [1, 41].

The confidence interval for s is obtained after rearranging this inequality as follows:

$$\frac{2 \times M \times accuracy + z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4M \times accuracy - 4M \times acccracy^2}}{2\left(M + z^2_{\alpha/2}\right)} \quad (9)$$

V. RESULTS AND ANALYSIS

In this section, we present empirical data that demonstrate the efficacy and performance of the proposed approach, as well as its comparison with other classifiers such as a naive Bayes model, SVM, Random Forest, KNN, and Logistic Regression. We compare the proposed approach's performance (95% of data) to that of others. Table 2 displays the accuracy, precision, recall, F1-score, and AUC of techniques chosen at random (100% of the data) on White Wine datasets.

5.1 Overall Performance Analysis

On using machine learning techniques NB, SVM, RF, KNN, and LR for white wine dataset (100%) with random sampling approach, the overall performance analysis is depicted in Table 2. Here, evaluation parameters accuracy, precision, recall and F1-score`for RF are greater than other methods.

Table 2 Performance Evaluation of Techniques for the White Wine Dataset With 100% Data

Techniques					
Measures	NB	SVM	RF	KNN	LR
Accuracy	0.73	0.73	**0.83**	0.75	0.72
Precision	0.72	0.72	**0.83**	0.75	0.71
Recall	0.73	0.73	**0.83**	0.75	0.72
F1-score	0.72	0.72	**0.83**	0.75	0.71

5.2 Performance Metrics With 95% Data with SC & Random Sampling

The graph plots in Fig 4 illustrates that the proposed approach sampling from cluster SC is superior to Random sampling for the 2-class white wine dataset. The SC gives the highest accuracy 75%, 76%, 84%, 76%, & 77%; highest precision 75%, 75%, 84%, 75% & 76%; highest recall 75%, 76%, 84%, 76% & 77%; and highest F1-score 73%, 75%, 83%, 75% & 76% for NB, SVM, RF, KNN(k=5) & LR respectively for the white wine data subset. In the meanwhile, the Random sampling technique (RS) gives the accuracy 74%, 74%, 83%, 75%, & 74%; the precision 73%, 73%, 83%, 75%, & 73%; the recall 74%, 74%, 83%, 75%, & 74%; the F1-score 72%, 74%, 83%, 75%, and 73% for NB, SVM, RF, KNN (k=5) & LR respectively for the white wine data subset. The ROC (Receiver Operating Characteristic) Curves for 95% of data subsets are presented in Fig 5. Here, Area under Curve (AUC) for machine learning techniques, NB, SVM, and LR using proposed approach SFC is increased by 1%, 0.6%, & 2.7% respectively whereas for the techniques RF & KNN using RS is incremented by 0.1% & 0.7% respectively. The corresponding AUC for 95% data for the RS technique for specified machine learning techniques are 0.660, 0.691, 0.796, 0.713, & 0.684 respectively. Whereas, corresponding AUC of the SC for 95% data of white wine dataset using machine learning techniques NB, SVM, RF, KNN, & LR are 0.670, 0.697, 0.795, 0.706, & 0.711.

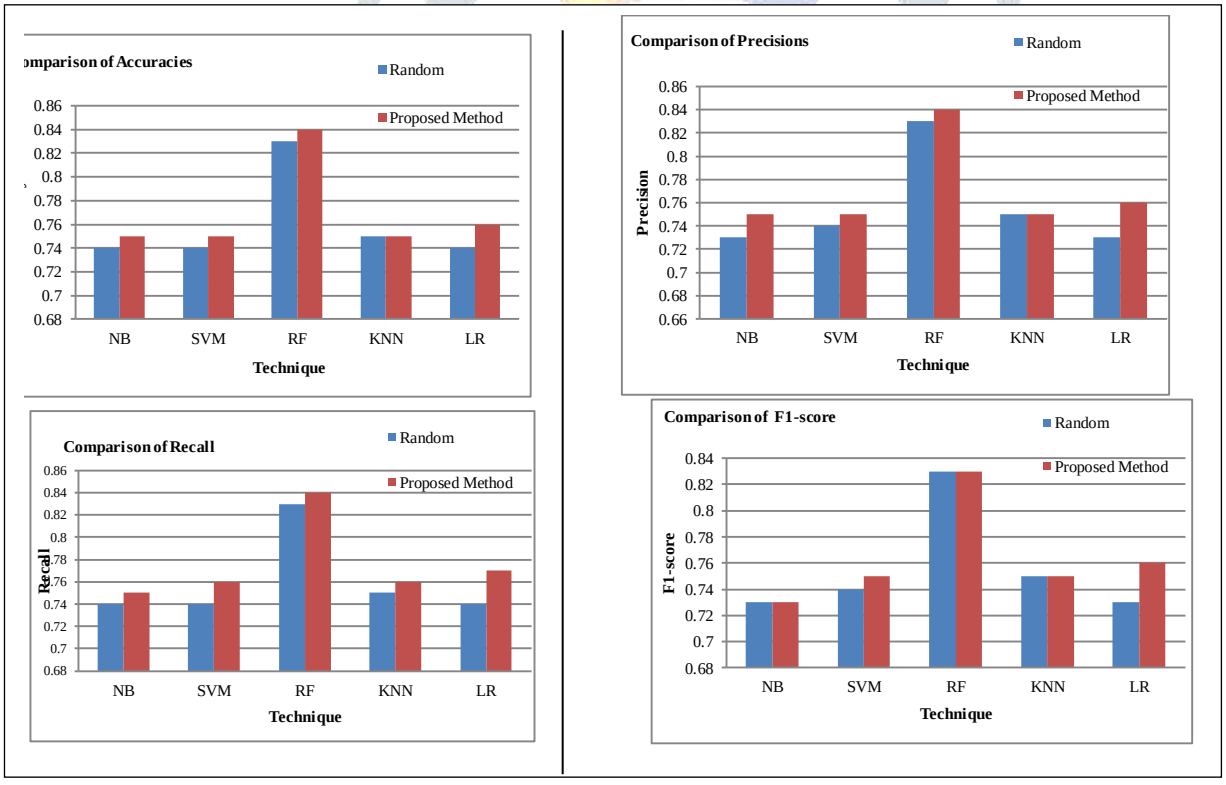

Fig. 4 Overall Performance analysis for data subsets (95% data)

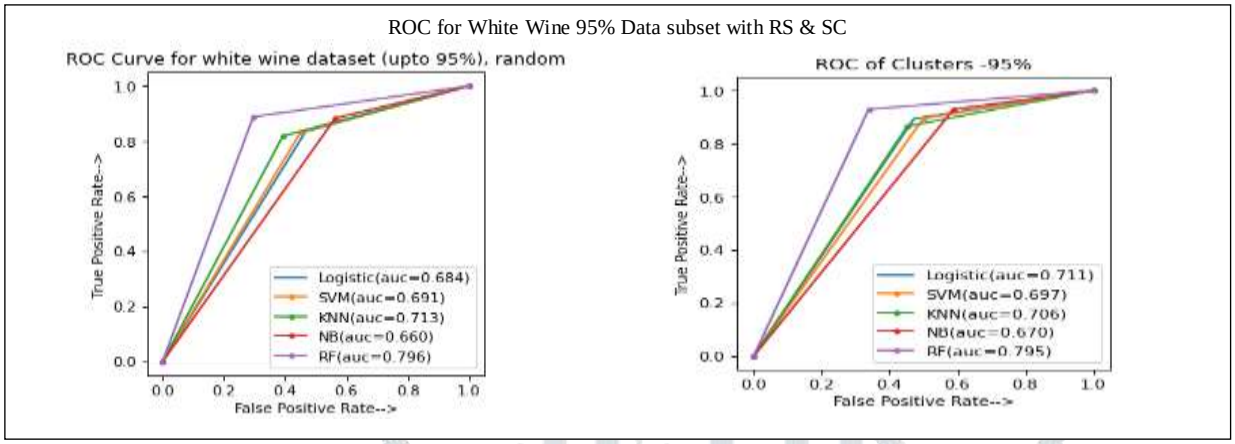

Fig. 5 ROC Curve for the white wine (95%) data.

5.3 Performance metrics with 90% SC & Random Sampling Data

The graph plots in Fig 6 show that the proposed approach sampling from cluster SC outperforms random sampling for the datasets with 2- classes. For the white wine data subset, the SC provides the highest accuracy (74%, 74%, 82%, 77%, & 75%), highest precision (74%, 74%, 82%, 76% & 74%), highest recall (74%, 74%, 82%, 77% & 75%), and highest F1-score (72%, 73%, 82%, 76% & 73%), respectively for NB, SVM, RF, KNN(k=5) & LR respectively for the white wine data subset. In the meanwhile, the Random sampling technique RS gives the accuracy (71%, 73%, 81%, 74%, &73%); the precision (70%, 72%, 80%, 74%, & 72%); the recall (71%, 73%, 81%, 74%, & 73%); the F1-score (69%, 72%, 80%, 74%, & 72%), respectively for NB, SVM, RF, KNN (k=5) & LR respectively for the white wine data subset.

The ROC (Receiver Operating Characteristic) Curves for 90% of data subsets are presented in Fig 7. Here, Area under Curve (AUC) for machine learning techniques, NB, RF, KNN, and LR using proposed approach SFC is increased by 1.4%, 0.4%, 1.4% & 0.3% respectively whereas for the technique SVM using RS is incremented by 0.5% respectively. The AUC for the SC of 90% data for specified machine learning techniques NB, SVM, RF, KNN and LR are 0.652, 0.667, 0.776, 0.715, & 0.676 respectively.

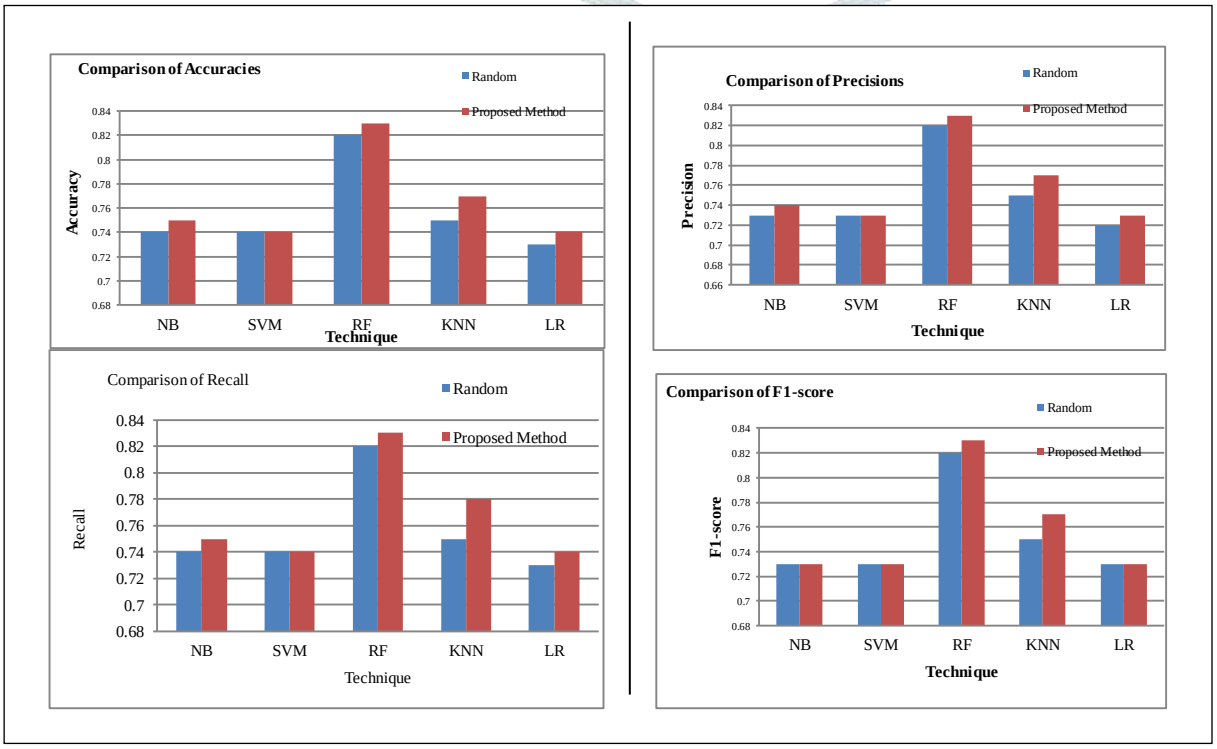

Fig 6 overall Performance analyses for data subsets (90% data)

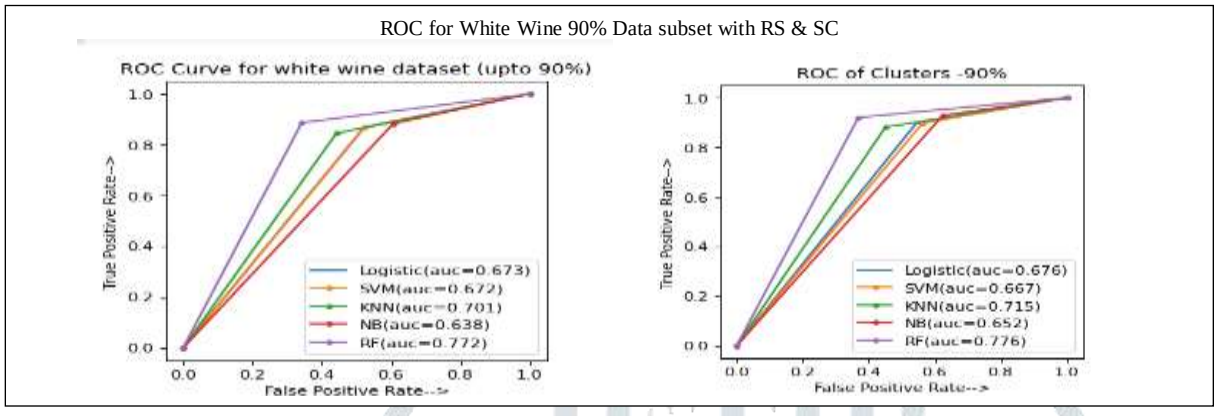

Fig. 7 ROC Curve for the white wine (90%) data.

5.4 Performance metrics with 85% SC & Random Sampling Data

It is evident from the graph designs in Fig. 8 that the proposed approach sampling from clusters SC outperforms random sampling for the datasets with 2- classes. For the white wine data subset, the SC provides the highest accuracy (74%, 74%, 82%, 77%, & 75%), precision (74%, 74%, 82%, 76% & 74%), recall (74%, 74%, 82%, 77% & 75%), and F1-score (72%, 73%, 82%, 76% & 73%), respectively.

In the meanwhile, the Random sampling method RS gives the accuracy (74%, 74%, 82%, 75%, &73%); the precision (73%, 73%, 82%, 75%, & 72%); the recall (74%,74%, 82%, 75%, & 73%); the F1-score (73%, 73%, 82%, 75%, & 73%), respectively for NB, SVM, RF, KNN (k=5) & LR respectively for the white wine data subset.

The ROC (Receiver Operating Characteristic) Curves for 85% of data subsets are presented in Fig 9. Here, Area under Curve (AUC) for machine learning techniques, NB, KNN, and LR using proposed approach SC is increased by 0.5%, 1.8%, & 0.3% respectively whereas for the technique SVM & RF using RS is incremented by 1.2% & 0.4% respectively.

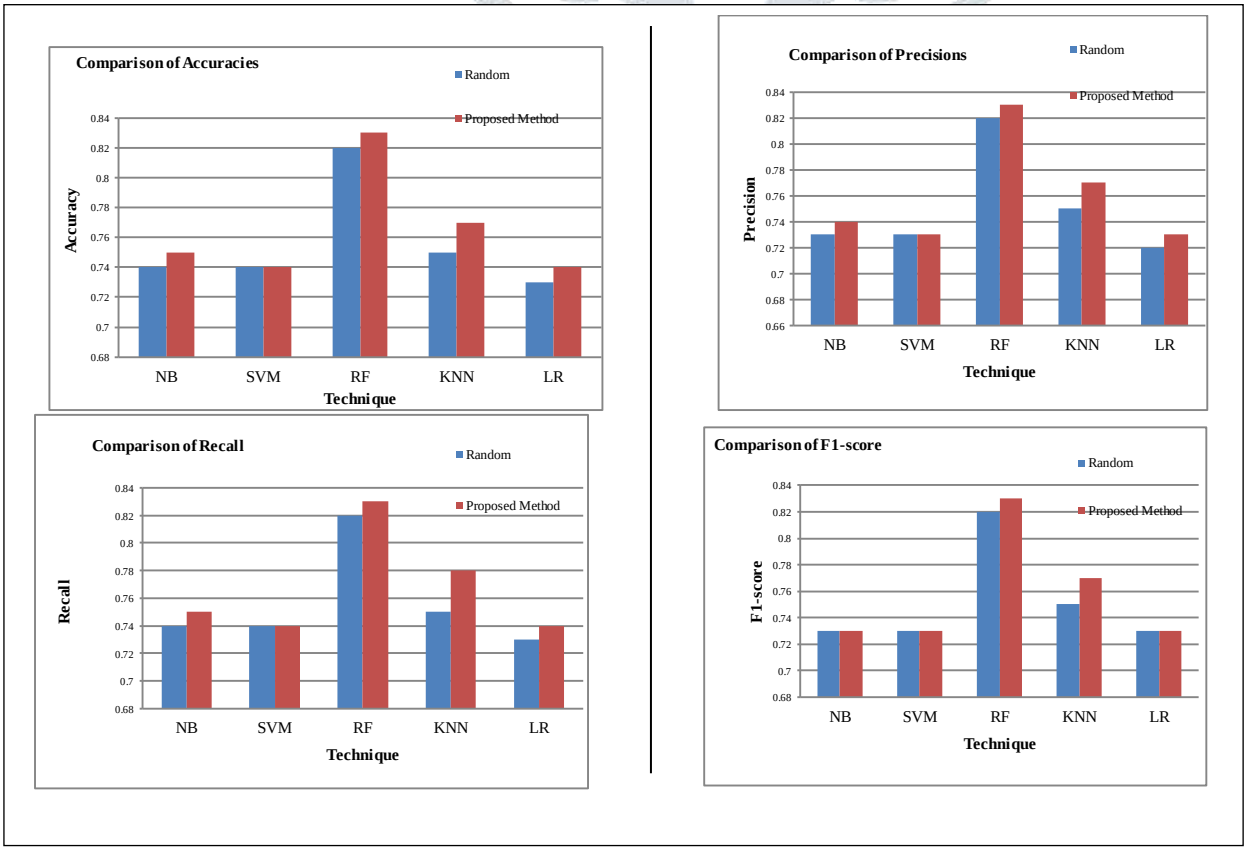

Fig. 8 Overall Performance analysis for data subsets

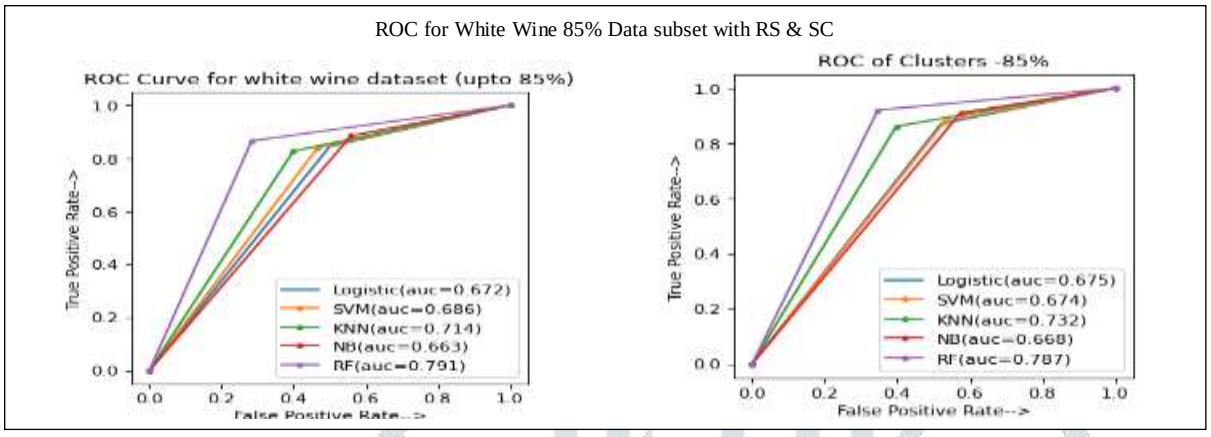

Fig. 9 ROC Curve for white wine (85%) data

5.5 Performance metrics with 80% SC & Random Sampling Data

The graph plots in Fig. 10 demonstrate that the proposed approach sampling from clusters SC is superior to Random Sampling for the 2-class white wine datasets. The SC gives the highest accuracy 75%, 74%, 81%, 74%, & 74%; highest recall 75%, 74%, 81%, 74% & 74% ; and highest F1-score 72%, 73%, 80%, 73% & 78% for NB, SVM, RF, KNN(k=5) & LR respectively for the white wine data subset.

In the meanwhile, the Random sampling method RS gives the accuracy (72%, 73%, 80%, 74%, &73%); the precision (71%, 72%, 80%, 73%, & 72%); the recall (72%, 73%, 80%, 74%, & 73%); the F1-score (71%,72%, 80%, 73%, & 73%), respectively for NB, SVM, RF, KNN (k=5) & LR respectively for the white wine data subset.

The ROC (Receiver Operating Characteristic) Curves for 80% of data subsets are presented in Fig 11. Here, Area under Curve (AUC) for machine learning techniques, NB using proposed approach SC is increased by 0.8% respectively whereas for the techniques SVM, RF, KNN, & LR using RS are incremented by 1.3%, 1.8%, 0.7%, & 1.4% respectively.

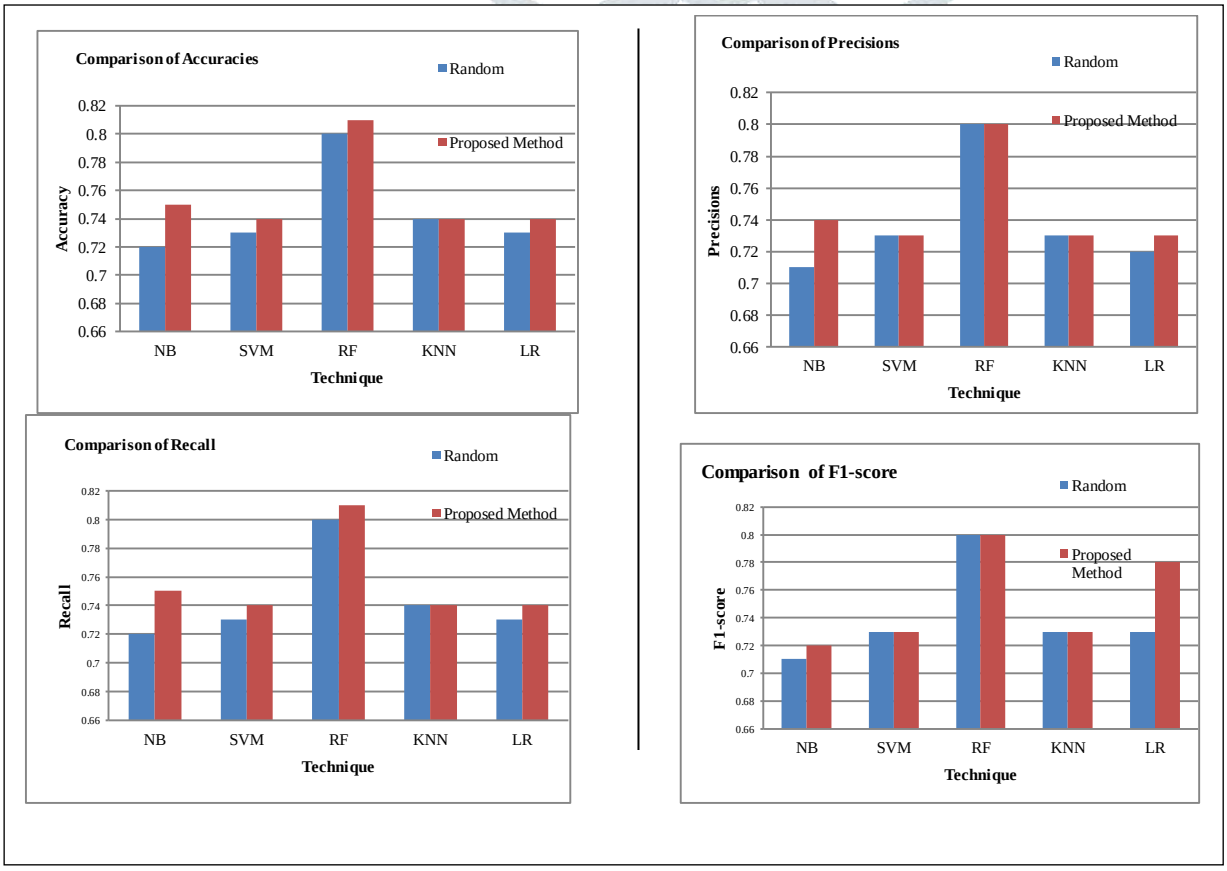

Fig. 10 Overall Performance analysis for data subsets

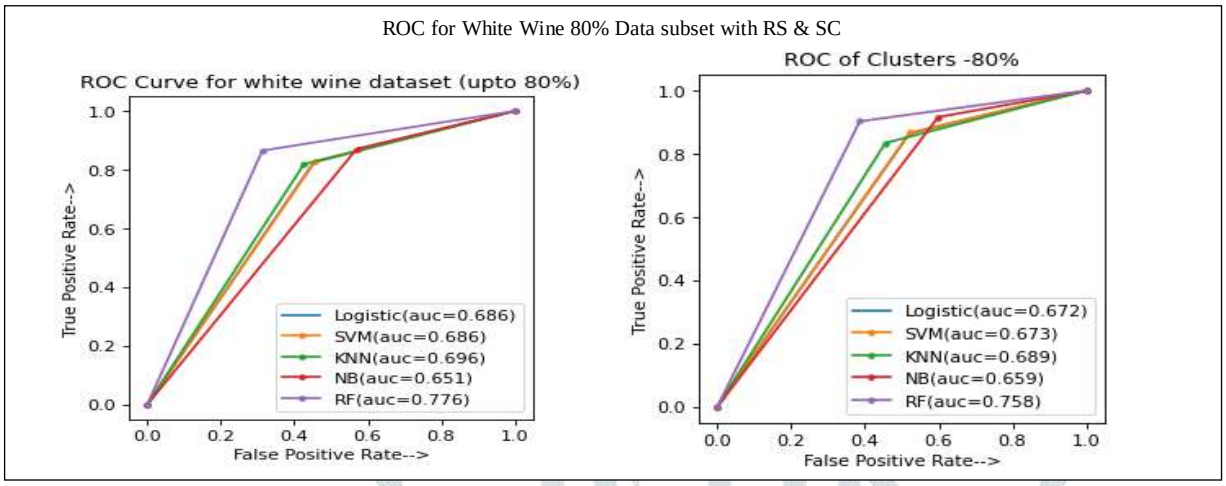

Fig. 11 ROC Curve for white wine (80%) data subset

5.6 MSE of SC & RS

The graph plots in Fig 12 demonstrate that MSE of NB, SVM, RF, KNN, and LR for Random sampling RS approach is larger than SC approach with 0.013, 0.003, 0.007, 0.006, and 0.31 respectively on using 95% data of given dataset.

Similarly, for 90% data, MSE of NB, SVM, RF, KNN, and LR for Random sampling RS approach is larger than SC approach with 0.034, 0.015 0.018, 0.027, and 0.018 respectively on using 90% data of given dataset. Similarly, for 85% data, MSE of NB, SVM, RF, KNN, and LR for RS approach is larger than SC approach with 0.007, 0.002, 0.015, 0.022, and 0.008 respectively on using 80% data of given dataset.

Similarly, for 80% data, MSE of NB, SVM, RF, KNN, and LR for RS approach is larger than SC approach with 0.024, 0.007, 0.002, 0.003, and 0.017 respectively on using 80% data of given dataset.

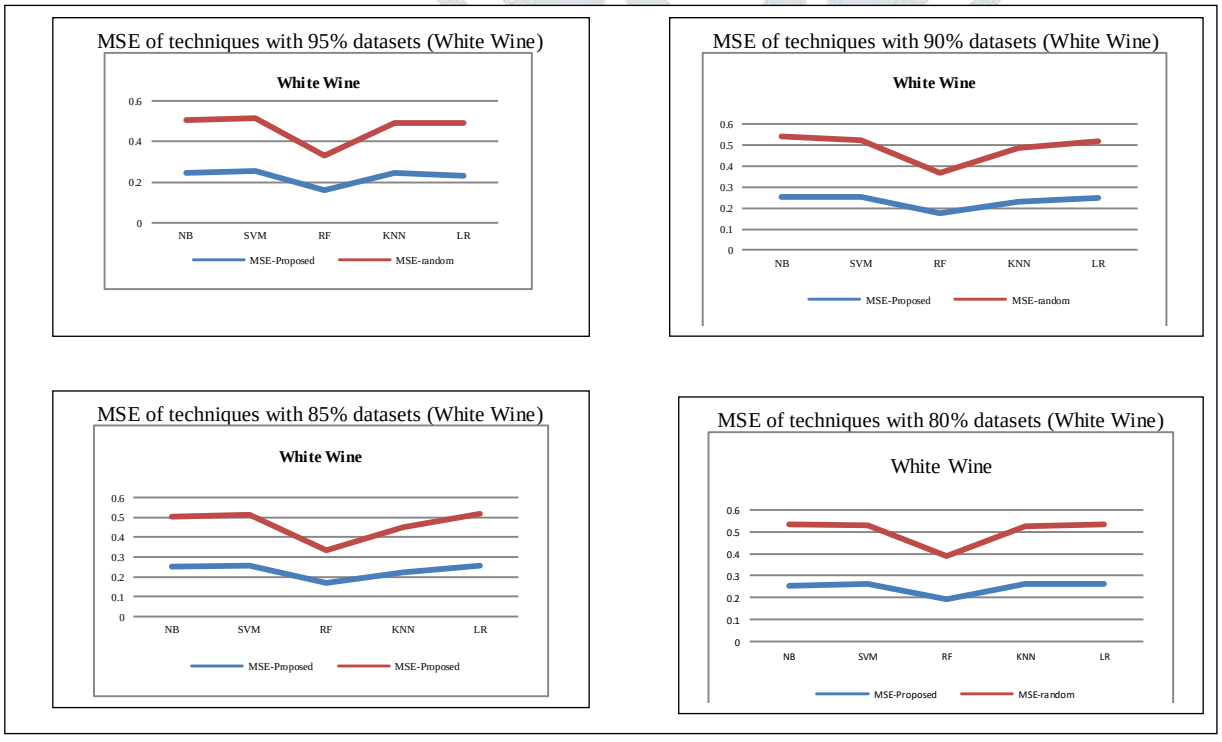

Fig. 12 MSE of Proposed method SC & RS

5.7 Statistical Evaluation

Using the accuracy, we also calculate the 95% confidence interval for each classifier. The techniques essentially provide an upper and lower bound for accuracy, representing all possible values for accuracy. A confidence interval (CI) is a statistical interval that describes the level of uncertainty in a given estimate. Along with a lower and upper bound, it also offers likelihood. A CI generally implies a minimal margin of error. This range is useful for estimating a model's potential. Standard calculations

show that they are 95%, 98%, and 99% respectively. A 95% CI signifies that 95% of studies will lie within the range, however 5% will not.

The calculated results of various experiments using the White Wine dataset with 95% confidence intervals are shown in Table 3. From this table we see that the accuracy values for the SC approach vary, as N increases, CI is tighter. At 50% labeled data, for Naïve Bayes is 0.74± 0.02, for SVM is 0.74 ± 0.02, for RF is 0.81 ± 0.01, for KNN is 0.73 ± 0.02 and for LR is 0.74 ± 0.02. This implies that the proposed model's efficiency is likely to be between 72% and 76% for NB, 72% to 76% for SVM, 80% to 82% for RF, 71% to 75% for KNN, and 72% to 76% for LR technique with 95% confidence.

Table 3 Confidence Interval for White Wine Data Subset (95% Data)

	N %				
Techniques	10%	20%	30%	40%	50%
NB	0.74 ± 0.04	0.75 ± 0.03	0.74 ± 0.02	0.73 ± 0.02	0.74 ± 0.02
SVM	0.75 ± 0.04	0.76 ± 0.03	0.75 ± 0.02	0.74 ± 0.02	0.74 ± 0.02
RF	0.85 ± 0.03	0.84 ± 0.02	0.83 ± 0.02	0.82 ± 0.02	0.81 ± 0.01
KNN	0.76 ± 0.04	0.76 ± 0.01	0.74 ± 0.02	0.74 ± 0.02	0.73 ± 0.02
LR	0.76 ± 0.04	0.77 ± 0.03	0.75 ± 0.02	0.75 0.02	0.74 ±0.02

5.8 SC VS RS: Comparative Performance Analysis

Table 4 shows the accuracy of the proposed technique SC as well as the accuracy of the random sampling technique RS for the specified classifiers NB, SVM, RF, KNN, and LR with 100%, 95%, 90%, 85%, and 80% data of White Wine. In this table, we have observed an overall improvement is **up to 3%** using our proposed technique SC. It is also observed that the accuracy of the RF for both the proposed technique SC and random sampling RS is the highest among all specified classifiers used in this paper.

Table 4 Accuracy of White Wine Data with Techniques SC & RS

	Techniques									
% of data	Naïve Bayes		SVM		Random Forest		KNN		Logistic Regression	
	SC	RS	SC	RS	SC	RS	SC	RS	SC	RS
100%	0.73	0.73	0.73	0.73	0.83	0.83	0.75	0.75	0.72	0.72
95%	**0.75**	0.74	**0.76**	0.74	**0.84**	0.83	**0.76**	0.75	**0.77**	0.74
90%	**0.74**	0.71	**0.74**	0.73	**0.82**	0.81	**0.77**	0.74	**0.75**	0.73
85%	**0.75**	0.74	0.74	0.74	**0.83**	0.82	**0.78**	0.75	**0.74**	0.73
80%	**0.75**	0.72	**0.74**	0.73	**0.81**	0.80	0.74	0.74	**0.74**	0.73

Table 5 displays the precision of the proposed approach SC as well as the precision of the random sampling technique RS for the specified classifiers NB, SVM, RF, KNN, and LR with 100%, 95%, 90%, 85%, and 80% White Wine data, respectively. We observed an overall improvement of **up to 3%** using our proposed approach SC in this table.

The precision of the RF for both the proposed approach SC and random sampling RS was observed to be the highest among all techniques.

Table 5 Precision of White Wine Data with Techniques SC & RS

% of data	*Techniques*									
	Naïve Bayes		SVM		Random Forest		KNN		Logistic Regression	
	SC	RS	SC	RS	SC	RS	SC	RS	SC	RS
100%	0.72	0.72	0.72	0.72	0.83	0.83	0.75	0.75	0.71	0.71
95%	**0.75**	0.73	**0.75**	0.74	**0.84**	0.83	0.75	0.75	**0.76**	0.73
90%	**0.74**	0.70	**0.74**	0.72	**0.82**	0.80	**0.76**	0.74	**0.74**	0.72
85%	**0.74**	0.73	0.73	0.73	**0.83**	0.82	**0.77**	0.75	**0.73**	0.72
80%	**0.74**	0.71	0.73	0.73	0.80	0.80	0.73	0.73	**0.73**	0.72

Table 6 shows the recall for the specified classifiers NB, SVM, RF, KNN, and LR with 100%, 95%, 90%, 85%, and 80% White Wine data, as well as the recall for the proposed approach SC and RS. In this table, we found an overall improvement of **up to 3%** using our suggested method SC. The proposed approach SC and random sampling RS were both found to have the highest recall of the RF of all the techniques. Table 7 displays the F1-score of the proposed approach SC as well as the F1-score of the random sampling method RS for the specified classifiers NB, SVM, RF, KNN, and LR with 100%, 95%, 90%, 85%, and 80% White Wine data, respectively. We observed an overall improvement of **up to 3%** using our proposed method SC in this table. The F1-score of the RF for both the proposed method SC and random sampling RS was observed to be the highest among all methods.

Table 6 Recall of White Wine Data with Techniques SC & RS

Techniques										
% of data	Naïve Bayes		SVM		Random Forest		KNN		Logistic Regression	
	SC	RS	SC	RS	SC	RS	SC	RS	SC	RS
100%	0.73	0.73	0.73	0.73	0.83	0.83	0.75	0.75	0.72	0.72
95%	**0.75**	0.74	**0.76**	0.74	**0.84**	0.83	**0.76**	0.75	**0.77**	0.74
90%	**0.74**	0.71	**0.74**	0.73	**0.82**	0.81	**0.77**	0.74	**0.75**	0.73
85%	**0.75**	0.74	0.74	0.74	**0.83**	0.82	**0.78**	0.75	**0.74**	0.73
80%	**0.75**	0.72	0.74	0.73	**0.81**	0.80	0.74	0.74	**0.74**	0.73

Table 7 F1-score of White Wine Data with Techniques SC & RS

Techniques										
% of data	Naïve Bayes		SVM		Random Forest		KNN		Logistic Regression	
	SC	RS	SC	RS	SC	RS	SC	RS	SC	RS
100%	0.72	0.72	0.72	0.72	0.83	0.83	0.75	0.75	0.71	0.71
95%	0.73	0.73	**0.75**	0.74	0.83	0.83	0.75	0.75	**0.76**	0.73
90%	**0.72**	0.69	**0.73**	0.72	**0.82**	0.80	**0.76**	0.74	**0.73**	0.72
85%	0.73	0.73	0.73	0.74	**0.83**	0.82	**0.77**	0.75	0.73	0.73
80%	**0.72**	0.71	0.73	0.73	0.80	0.80	0.73	0.73	**0.78**	0.73

VI. CONCLUSION

Since managing enormous datasets in the real world is difficult, it is necessary to minimize the size of the data set, so that the accuracy of the original dataset is no longer impacted. In this study, K means Clustering technique is used to create clusters, since clustering gives a particular shape from which training samples are selected, and these clusters are used for the classification of white wine data. To the best of our knowledge, no one has before created a wine categorization based on labeled data using SC. Our proposed approach SC applies to the datasets to generate clusters first, and then 95% of the data from each cluster is extracted and merged to generate reduced datasets for use in classification models such as NB, SVM, RF, KNN, and LR. In the same way, 90% of the data from each cluster are taken and merged for further processing. The same process is repeated with 85% and 80% of the data from clusters. Performances for each classifier are evaluated. In the meanwhile, on the other side, we have worked with 95% of data taken randomly from the white wine dataset in a random sampling technique and compared the performances with our approach SC. The same process is repeated with 90% data, 85% data, and 80% data of the White Wine dataset. It has been shown that the approach we propose, SC, offers more accuracy than random sampling (RS) and it is also observed that RF provides higher accuracy than SVM, LR, KNN, and NB.

MSE for the proposed approach SC has less variability than the random sampling proximity. Figures 4, 6, 8, and 10 illustrate the white wine's accuracy, precision, recall, and F1-score values; SC improves performances significantly when 95%, 90%, 85%, and 80% data are used. The amount of labeled data utilized for supervised training is raised in order to strengthen the SC's performance. The ROC curves in Figures 5, 7, 9, and 11 also show that SC performs better results.

Furthermore, the results of the SC are statistically significant for the accuracy of the specified methods as depicted in Table 4 on increasing the number of test data (N), confidence intervals (CI) are closer, varying from 0.72 to 0.76 for NB, 0.73 to 0.79 for SVM, 0.82 to 0.86 for RF, 0.75 to 0.77 for KNN and 0.74 to 0.80 for LR. After analysis, it is found that reduction of the size of dataset using proposed approach SC gives better results than the original size of the dataset. Also, Random Forest technique is the most successful classification method for RS and SC.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Data Availability The dataset is taken from UCI Machine Repository [40].

REFERENCES

[1] Tan, P.N., Steinbach, M., Karpatne, A., & Kumar, V. (2022). *Introduction to Data Mining.* 2nd ed., Pearson Publications.

[2] Dunham, M.H. (2013). *Data Mining Introductory and Advanced Topics*. 14th ed., Pearson Education.

[3] Han, J., Kamber, M., & Pei, J. (2008). *Data Mining Concepts & Techniques*. 3rd ed., Morgan Kaufmann Publishers, ISBN: 978-93-80931-91-3.

[4] Ahsaan, S.U., Kaur, H., Mourya, A.K., & Naaz, S. (2022). A Hybrid Support Vector Machine Algorithm for Big Data Heterogeneity Using Machine Learning. (*MDPI*), Symmetry 2022, 14, 2344. https://doi.org/10.3390/sym14112344

[5] Sharma, N. (2018). Quality Prediction of Red Wine based on Different Features Sets Using Machine Learning Techniques. *International Journal of Science and Research (IJSR),* ISSN: 2319-7064, Research Gate Impact Factor.

[6] Parvathy, V.A., & Joseph, J. (2021). Comparative Analysis of Classification Algorithm for Predicting Wine Quality Using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*. ISSN: 2582-5208, Volume: 03/Issue:09/ September.

[7] Akanbi, O.D., Faloni, T.M., & Olaniyi, S. (2022). Prediction of Wine Quality: Comparing Machine Learning Models in R Programming. *International Journal of Latest Technology in Engineering. Management & Applied Science (IJLTEMAS)*, Volume XI, Issue IX, ISSN 2278-2540.

[8] Yesim, E., & Atasoy A. (2016). The classification of White Wine and Red Wine According to Their Physicochemical Qualities. *International Journal of Intelligent Systems and Applications in Engineering (IJISAE),* 4(Special Issue), 23-26, ISSN: 2147-6799.

[9] Kumar, S., Agarwal, K., & Mandan, N. (2020). Red Wine Quality Prediction Using Machine Learning Techniques. Conference paper, *Research Gate*.

[10] Korade, N., & Salunke, M.,: Identification of appropriate Machine Learning Algorithm to Predict Wine Quality. *International Journal of Scientific Research in Engineering and Management (IJSREM),* Volume: 05 Issue: 05,ISSN: 2582-3930. (2021)

[11] Chen, B., Rhodes, C., Crawford, & A., Hambuchen, L. (2014). Wine informatics: applying data mining on wine sensory reviews processed by the computational wine wheels. *IEEE International Conference on Data Mining Workshop*, pp. 142-149.

[12] Guia, M., Silva, & R.R., Bernardino, J. (2019). Comparison of NB, SVM, DT and RF of Sentiment Analysis. Proceedings of the *11th International Joint Conference on Knowledge Discovery Knowledge Engineering and Knowledge Management (JC3K 2019)* pages 525-531, ISSN-978-989-756-382-7.

[13] Mishra, A., Jha R., & Bhattacharjee, V. (2023). SSCL Net: A Self-Supervised Contrastive Loss-Based Pre-Trained Network for Brain MRI Classification. *IEEE Access*.

[14] Rahmadani, S., Dongoran, A., Zarlis M., & Zakarias. (2017). Comparison of NB and DT on Feature Selection Using Genetic Algorithm for Classification Problem. *2nd International Conf. on Computing and Applied Informatics*, IOP Conf. Series: Journal of Physics: 978 (2018) 012087.

[15] Wibowo, A.H., & Oesman, T.I. (2019). The comparative analysis on the accuracy of k-NN, NB, and DT Algorithms in predicting crimes and criminal actions in Sleman Regency. iCAST-ES 2019, IOP Publishing, *Journal of Physics: Conference Series*, 1450 (2020) 012076.

[16] Patil, T.R., & Sherekar, S.S. (2013). Performance Analysis of NB and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science And Applications*, Vol. 6, No.2.

[17] Untoro M.C., Praseptiawan, M., Widianingsih, M., Ashari, I.F. A. Afriansyah., & Oktafianto. (2019). Evaluation of Decision Tree, K-NN, NB and SVM with MWMOTE on UCI Dataset. ICComSET 2019. IOP Publishing, *Journal of Physics: Conference Series*, 1477 - 032005.

[18] Karthika, S., & Sairam, N. (2015). A Naïve Bayesian Classifier for Educational Qualification. *Indian Journal of Science and Technology*, Vol 8(16). ISSN (Online) : 0974-5645. DOI: 10.17485/ijst/2015/v8i16/62055.

[19] Sharma, N. (2016). Classification Using Naïve Bayes – A Survey. *International Journal of Engineering Science Invention Research & Development*; Vol. II Issue VIII.

[20] Uddin, S., Khan, A., Md. Hossain, E., & Md. Moni A. (2019). Comparing Different supervised machine learning algorithms for disease prediction. *BMC Medical Informations and Decision Making* 19-28.

[21] Sheth, V., Tripathi, U., & Sharma, A. (2020). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, Volume 215, pages 422-431.

[22] Noi, P.T., & Kappas, M. (2018). Comparison of RF, K-Nearest Neighbour, and Support Vector Machine classifiers for Land cover Classification Using Sentinet-2 Imagery. *Sensors (Basel)* , PMC5796274, 18(1):18.

[23] Iqbal S.M.H.S., Jahan N., Moni, A.S., & Khatun, M. (2022). An Effective Analytics and Performance Measurement of Different Machine Learning Algorithms for Predicting Heart Diseases. *International Journal of Advanced Computer Science and Applications*, Vol 13, No. 2.

[24] Bhardwaj, P., Tiwari, P., Olejar Jr K., Parr W., & Kulasiri, D. (2022). A Machine Learning application in wine quality prediction. *Machine Learning with Applications*, Vol 8, 100261.

[25] Mabayoje, M.A., Balogun, A.O., Salihu, S., & Oladepupo, K.R. (2015). Comparative analysis of Selected Supervised Classification Algorithms. *African Journal of Computing & ICT*, Vol 8, No. 3(2), ISSN 2006-1781. (IEEE).

[26] Yuvali, M., Yaman, B., & Tosun, O. (2022). Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. *Mathematics 2022*, Vol 10,311. https://doi.org/10.3390/math10030311

[27] Reddy, R.V.K., & Babu, V.R. (2018). A Review on Classification Techniques in Machine Learning. *International Journal of Advance Research in Science and Engineering (IJARSE)*, Volume No. 07, Special Issue No. 03.

[28] Grewal, P., Sharma, P., Rathee, A., & Gupta, S. (2022). Comparative Analysis of Machine Learning Models. *EPRA International Journal of Research and Development (IJRD)*, Volume: 7| Issue:6, ISSN: 2455-7838(online).

[29] Tan, H. (2021). Machine Learning Algorithm for Classificatio. *International Conference on Big Data and Intelligent Algorithms (BDIA 2021), Journal of Physics: Conference Series*.

[30] Grgi, V., Music, D., & Babovic, E. (2021). Model for predicting heart failure using Random Forest and Logistic Regression algorithms. *IOP Conference Series: Materials Science and Engineering* 1208012039.

[31] Cao, Y., Chen, H., & Lin, B. (2022). Wine Type Classification Using Random Forest. Highlights in Science, *Engineering, and Technology*, SDPIT2022, Volume 4.

[32] Couronne, R., Probst, P., & Boulestei, A.L. (2018). Random forest versus logistic regression: a large–scale benchmark experiment. et al. *BMC Bioinformatics*, 19:270.

[33] Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs. Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, Volume 1, Number 3, Article 9.

[34] Lingjun, H., Levine, R.A., Fan, J., Beemer, J., & Stronach, J. .(2018). Random Forest as a Predictive Analytics Alternative to Regression in *Institutional Research. ISSN 1531-7714*, Volume 23, Number 1.

[35] Tigga, O., Pal, J., & Mustafi, D. (2023). A Comparative Study of Multiple Linear Regression and KNNs using Machine Learning. *Fifth IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).*

[36] Itoo, F., Mittal, M., & Singh, S. (2020). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology.*

[37] Boateng, E.Y., Otto, J., & Abaye, D.A. (2020). Basic Tenets of Classification Algorithms K-Nearest Neighbour, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analytics and Information Processing*, Vol 8 No. 4.

[38] Kumari, A.D., Kumar, J.P., & Prakash,V.S. (2020). Supervised Learning Algorithms: A Comparison. *Kristu Jayanti Journal of Computational Sciences*, Vol.1, Issue 1, pp. 01-12.

[39] Khire, S., Ganorkar, P., Apastamb, A., & Panicker, S. (2020). Investigating the Impact of Data Analysis and Classification on Parametric and Nonparametric Machine Learning Techniques: A Proof of Concept. *Computer Networks and Inventive Communication Techniques*, Proceedings of Third ICCNCT.

[40] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[41] Tigga, O., Pal, J., & Mustafi, D. (2022). A Comparative Study of Rule-Based Classifier and DT in Machine Learning. *4th International Conference on Soft Computing and its Engineering Applications (ICSOFTCOMP).*

[42] Pal, J., Mustafi, D., & Tigga, O. (2022). Using Hierarchical Fuzzy Rule-Based System to Predict Software Quality. *2nd International Conference on Nano Electronics, Machine Learning, Internet of Things & Computing Systems (NMIC).*