SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I7P101

Volume 12 Issue 7, 1-12, July 2025

Original Article

A Semi-Supervised Ensemble Classification with Data Augmentation to Reduce Imbalanced Class Distribution in Big Data

S. Sindhu[1], S. Veni[2]

[1,2]*Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.*

Corresponding Author : sssindhu784@gmail.com

Received: 01 May 2025 Revised: 03 June 2025 Accepted: 02 July 2025 Published: 31 July 2025

Abstract *- With the age of digital information, managing, retrieving, and processing information in the form of big data has become the most substantial and difficult task. Though several tools are available for data analysis, they may not perform well with high-dimensional data. Machine learning algorithms play a major role in acquiring useful knowledge from this massive data. Consequently, a better technique that is extremely fast in handling a huge volume of data with numerous dimensions and processing ability in a distributed setting is indispensable. As the data has been gathered from various sources, data preprocessing plays a vital role in transforming it into a suitable form for acquiring useful insights. Despite the data being pre-processed, class imbalance remains a significant concern that must be addressed to enhance the outcomes. This work proposes a semi-supervised ensemble classifier utilizing data augmentation and the classroom rebalancing sampling to mitigate uneven class distribution in large datasets. The model employs ensemble techniques with two heterogeneous classifiers: extreme gradient boosting as a base learner and random forest for pseudo-labelling and classification. The Hadoop framework implements the model, which is suitable and effective for very large datasets. Following the execution of the experimental evaluation for the suggested framework, the evaluation metrics of the results indicate that the approach achieves an average accuracy of 84.3% over 16 datasets, reflecting an increase of around 17.65% in prediction performance.*

Keywords *- Class imbalance, Data augmentation, Ensemble learning, Extreme gradient boosting, Semi-supervised learning.*

1. Introduction

Big data has been used extensively in information technology in the past three decades, particularly since 1990. The word "big data" describes large amounts of data that are extremely large or complex to handle or analyze by traditional data processing applications [1]. However, due to the internet evolution at the beginning of the digital era in 2002, analogue storage fading took place, and from then on, the non-linear progress of digital storage on global information has been perceived [2]. However, in recent years, the term big data has been used to refer to predictive or advanced analysis to extract interesting knowledge or valuable insight from previous data and to visualise future trends [3]. Thus, big data is initially identified based on the three key elements: volume, velocity, and variety, whereas later veracity and value are also included [4, 5].

Machine learning is generally used to process, analyze and predict future insights from the given input dataset. However, due to various challenges, the techniques or prediction algorithms used for small and medium datasets cannot be directly utilized for big data [6]. Since big data is collected from different sources, the data will be in different formats, including structured, semi-structured and unstructured instances with many attributes. So the algorithms pose a complex challenge for predicting information [7]. Though many evolutionary learning algorithms are anticipated in the literature for performing various tasks, most of them do not address the issue related to scalability [8]. Thus, it is inevitable to reformulate the machine learning algorithms that scale to meet big data requirements [9, 10].

To identify the valuable knowledge, various steps including pre-processing, feature selection, classification, prediction and visualization must be efficient for effective results since each step contributes to the accuracy [11]. Classification is a frequently and widely used procedure for various applications that may use supervised or semi-supervised learning models to predict class labels [12]. Recent studies have used Ensemble learners to increase classification performance and accuracy [13]. However, the main issue faced by the classification algorithms is the class imbalance that contains majority and minority classes, due to which classification results undergo a bias towards the majority class

[14, 15]. A wide range of solutions exist in solving an imbalanced class problem, such as resampling by 1) oversampling or increasing the minority samples [16], 2) downsampling or decreasing the majority samples [17, 18], class probabilities [19] and other variations [20].

One of the solutions for imbalanced class distribution that seeks less attention is resampling through the Semi-Supervised Learning (SSL) model. A literature survey on semi-supervised learning models shows the importance of various applications [21]. A semi-supervised learning model was proposed that uses a Support Vector Machine (SVM) to improve the learning performance [22]. A graph-based learning model was proposed that utilizes kernel Hilbert spaces and spectral models [23]. A semi-supervised model using deep learning networks on top of the Ladder network is designed for deep unsupervised learning [24] by combining the supervision models [25] - The Co-forest classifier. A semi-supervised learner was proposed to use the co-training paradigm with Random Forest (RF) to assess the power of label samples using undiagnosed samples [26].

SSL generally uses labeled and unlabeled data to create a balanced class and enhance the learning model's efficiency [27]. In SSL, Training samples, also known as pseudo-labeled samples, are used to identify the class labels for the test data once the model has been trained. The model is then retrained using labeled and pseudo-labeled samples for classification [28]. Thus, class rebalancing sampling is used to increase the minority class samples to increase classification accuracy [19].

One widely used big data paradigm is MapReduce, a distributed programming framework with two primary stages: the map phase, which splits and processes the data and the reduce phase, which accumulates results [29]. Thus, owing to the advantage of MapReduce for handling big data, this paper presents a semi-supervised ensemble classification model that makes use of XGBoost and RF classifiers with data augmentation using constrained class rebalancing to reduce the class imbalance in big data and to improve the classification results. The model uses weight-based aggregation with Bayesian information reward to predict the pseudo labels. Several experimental analyses are used to assess the model, and the findings indicate that the suggested approach provides better outcomes for big datasets.

The structure of the paper is as follows. Section 2 delineates the literature pertinent to the intended study. Section 3 details the suggested model, which includes a map-reduce framework and a detailed architecture of class-unbalanced datasets. Section 4 presents the experimental setup and the datasets used for the study. Section 5 discusses a detailed analysis of the acquired results and compares them with other existing models. Section 6 eventually ends the proposed study and identifies future research.

2. Related Work

This section provides an in-depth analysis of the literature pertinent to the proposed project, encompassing the classification of huge data, resampling techniques, and semi-supervised learning models. Data augmentation is widely used for balancing the samples of various classes, a significant field of study in improving the results of the underlying learning model. Regularized SVM was adopted for augmenting the data, which focuses on reducing the complexity of the model. However, it supports only binary classification [30]. In general, for performing data augmentation, various methods have been widely used, including Bayesian inference [31, 32], heuristic iterative approach [33], Markov Chain Monte Carlo (MCMC) inference, which is a sampling method that uses local information [34, 35]. Here, most methods do not support parallel computing [36] or distributed computing [9], which are essential for big data to minimize the time complexity.

Several algorithms were proposed in the literature for classification problems, irrespective of the applications. However, due to scalability issues, most learning techniques may not support big data. Only a few models, such as Extreme Learning Machines (ELM) [35], Extreme Gradient Boosting (XGBoost) [37] and SVM [38] support big data with improved performance. Nevertheless, big data classification in a distributed environment is still under research [39]. An Alternating Direction Method of Multipliers (ADMM) framework was proposed that splits the samples into subsamples, which are then processed in parallel in a distributed environment.

However, the model still needs improvement in accuracy and time complexity [40]. A fuzzy rule-based classification system for large quantities of data in a distributed environment was expected and demonstrated effectiveness in classification accuracy [41]. A similar model that utilizes MapReduce for classifying big data using linguistic fuzzy rules was proposed [29]. However, the above models employ fewer datasets for analyzing the performance.

Class imbalance is another severe problem that needs a persistent solution since it affects the results of any classifiers [42]. Semi-supervised learning algorithms are one such solution that gains attention in research by performing pseudo labeling to enhance the learning model's efficacy and balance the dataset's class distribution [43]. An unbiased semi-supervised learning model that utilizes XGBoost for an imbalanced dataset was proposed [44]. Unfortunately, the model was intended for image data and fails to examine the computational complexity. Moreover, class rebalancing using self-training was proposed that employs semi-supervised learning for effective results [20]. The method was proven to be effective, and thus it is used in the proposed model. An ensemble model that supports XGboost and transductive SVM (TSVM) was proposed [45]. However, the model fails to support datasets having a minimum number of labeled samples.

Though various methods are proposed related to the study, the classification of big data with imbalanced classes still needs to be focused on improving the efficiency of accuracy and computational time. Thus, an approach has been proposed that performs classification on big data with improved accuracy using class rebalancing and improves computational time with parallel computing in a distributed environment.

3. Proposed Pre-Processing Model

The proposed model aims to reduce the class imbalance due to the proportion variation between classes. The model utilizes a MapReduce framework to augment the minority class samples to improve the classification algorithm's efficiency. The semi-supervised model is utilized in the model, which utilizes two significant classifiers, such as XGBoost and RF, using bootstrapping to identify the pseudo labeling by Bayesian information reward-based weighted voting. Class rebalancing sampling is applied upon identifying the labels to augment the minority class samples. Each mapper applies XGBoost classifiers from the training datasets bootstrapped from labeled sets to classify the unlabeled datasets. The overall MapReduce framework for data augmentation-based semi-supervised learners is presented in Figure 1.

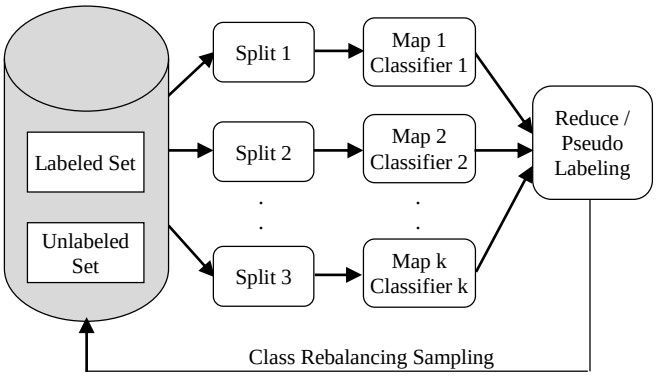


Fig. 1 A MapReduce framework for proposed data augmentation based classification

The proposed semi-supervised ensemble learning model has k classifiers. The first (k-1) classifiers are XGBoost, and the k^{th} classifier is the RF classifier, similar to the classification algorithm proposed in [45]. The predictions obtained from these sets of classifiers are used to predict the labels for the unlabeled sets and then to adjust the class imbalance. Here, the bootstrapped training sets are used by the XGBoost classifiers, whereas the RF classifier uses complete training sets. For k ensemble classifiers, the Bayesian information criterion that computes the probabilities of the prediction is used as a weight for the classifiers [46]. Thus, the k classifiers are employed to forecast class labels, significantly enhancing prediction accuracy. Finally, the class rebalancing sampling is applied to augment the data of the minority class, which helps in balancing the class proportions [20].

XGBoost is an enhanced boosting models that is particularly intended to advance the speed and performance of the classification model using decision trees [37]. It applies software and hardware optimization techniques to enhance performance with minimized resources and time. For creating the trees, the algorithm utilizes a similarity score in selecting and splitting nodes. The prediction is evaluated for each tree, and the residuals with the desired loss function are computed based on the similarity score calculated for selecting the nodes. Then the information gain is evaluated to construct the tree up to the specified level, and a regularization hyperparameter is used to prune and regularize the tree. The new residuals are computed from the previous trees. The (k-1) XGBoost classifiers used in the model apply bootstrapping or bagging on trained samples, thus helping to reduce computation complexity, decrease overfitting, and increase stability.

Upon training the model with XGBoost classifiers, the weights are assigned to the classifiers based on Bayesian information reward based on estimated probability p_i and prior probability p_i' where i represents the classes (1, 2, …, c) in the dataset as in Equations (1) and (2).

$$IR = \frac{\sum_i I_i}{c} \tag{1}$$

$$I_i = \begin{cases} I_i^+ = 1 - \dfrac{\log p_i}{\log p_i'} & for\ correct\ classifcation \\ I_i^- = 1 - \dfrac{\log(1-p_i)}{\log(1-p_i')} & for\ incorrect\ classifcation \end{cases} \tag{2}$$

Thus, the classification prediction results are organized by class labels, and the total weights of each classifier within each group are aggregated. Thus, the final results present the probability of predicted classes.

Similarly, an RF classifier is applied to the k^{th} classifier with the whole training set. This algorithm is most popular and used in many applications due to its improved prediction accuracy and ability to balance the classes. It constructs multiple decision trees for which it exploits bagging and feature randomness to generate independent trees to make a forest better than individual decision trees.

Here, the model is trained and the weight is computed for the classifier using Bayesian information reward. Upon implementing the RF classifier on the unlabeled dataset, it ends with a class prediction, which is then used for pseudo labeling along with the results provided by the XGBoost classifiers. Algorithm 1 explains the algorithm pseudocode for semi-supervised ensemble classifiers.

Algorithm 1: Algorithm for Semi-Supervised Ensemble Classifier

```
Input: Preprocessed Labeled set L and Unlabeled Sample Us
Output: k label prediction
Begin map_phase()
  For i from 1 to (k-1) do
    Split the labeled datasets into (k-1) random subsets with
      replacement
    Input the (k-1) subset of labeled samples to the (k-1)
      nodes
    Input all the labeled samples to the k^th node
  End For
  // Parallel-processing of nodes
  For all nodes n from 1 to k do
    If n<k then
      Train the model with XGBoost classifier
        For j from 1 to m do //number of iterations
          Compute initial tree, gradients and Hessians
          Fit the base learner
          Update the model
        End For
        Compute Bayesian information reward (IRn)
    Else
      Train the model and RF classifier
        For j from 1 to m do //number of trees
          Compute decision trees by selecting random
            samples and features
          Fit the base learner
          Aggregate the results
        End For
        Compute Bayesian information reward (IRn)
    End If
  End For
  For all nodes n from 1 to k do
    If n < k then
      Apply XGBoost classifier
      Predict the class label for Us
    Else
      Apply RF classifier
      Predict the class label for Us
    End If
  End For
  Return(Predicted Label, Information Reward)
End Procedure
```

Once the predicted labels and the information reward of the classifiers from the map nodes are received, the next step involves identifying the pseudo labeling from the ensemble classifiers using weight-based aggregation. Thus, the aggregate of the classifiers' weights relative to the classified class labels is calculated as delineated in Equation (3).

$$Psedo\_label(U_i) = \max_{C} \sum_{C} IR_k \tag{3}$$

Here, C represents the class variables that range from 1 to c. U_i represents the i^{th} instance of the unlabeled dataset, and k represents the classifiers. Thus, the classifiers with identical predicted labels are aggregated, and the information reward is totaled, with the label possessing the highest value designated as a final classification label for the unlabeled data. Thus, upon identifying the pseudo labeling, the class rebalancing sampling is applied in which rather than incorporating all the pseudo-labeled instances into the labeled dataset, only the selected samples are added based on the condition given below: the less the class label l in the labeled set L, the greater the number of unlabeled samples predicted as l that is incorporated into the labeled set L [20]. The unlabeled instances are included in the training set at the rate specified in Equation (4).

$$\vartheta_l = \frac{N_{C+1-l}}{N_{C_{major}}} \tag{4}$$

Here ϑ_l specifies the rate of inclusion of samples from class l, N_C indicates the classes in the datasets and $N_{C_{major}}$ specifies the sample count in the majority class C_{major}. Consider the class with 5 classes in which C_1 is the majority class with 100 samples (N_1) and C_5 is the minority class with 10 samples (N_5). Thus, the imbalanced ratio for the dataset can be given as 10 $\left(=\frac{N_1}{N_5}\right)$. Thus, the rate of inclusion of the samples from the minority class (C_5) is 1 $\left(=\frac{N_{5+1-5}}{N_1}\right)$ whereas the rate of inclusion of the samples from the majority class (C_1) is 0.1 $\left(\frac{N_{5+1-1}}{N_1}\right)$. This helps to augment the minority class instances, which even helps to increase prediction accuracy. The algorithm pseudocode for the pseudo labeling using weight aggregation and the class rebalancing sampling for balancing the imbalanced class distribution is presented in Algorithm 2.

Algorithm 2: Pseudo Labeling and Class Rebalancing Sampling

```
Input: Preprocessed Labeled set L and Unlabeled Sample Us, set
of k prediction with IR (Pk, IRk)
Output: Final prediction
Begin reduce_phase()
  For each class j in C do
    For i ranges from 1 to k do
      If Pi == Cj then
        Weightj = Weightj+ IRi
      End If
    End For
  End For
  //Pseudo labeling step
  For each class j in C do
    Select the maximum class weight for pseudo labeling Cw
     for Us
  End For
  //Class Rebalancing Sampling process
  For each class C in L do
    Count the instances count in each class as N
    Identify the majority class Cmajor
    If N(Cw) < N(Cmajor) then
      Include the predicted sample into the labeled sample
      Update the samples
    End If
  End For
End Procedure
```

Despite using RF and XGBoost classifiers to improve the classification result's accuracy, using a semi-supervised model to augment the data to reduce the bias caused by the class imbalance also highly helps produce improved performance. Similarly, the computational complexity and time complexity are greatly decreased by using MapReduce, which not only bootstraps but also completes the classification process in parallel [47]. The detailed working steps of the proposed model are shown in Figure 2.

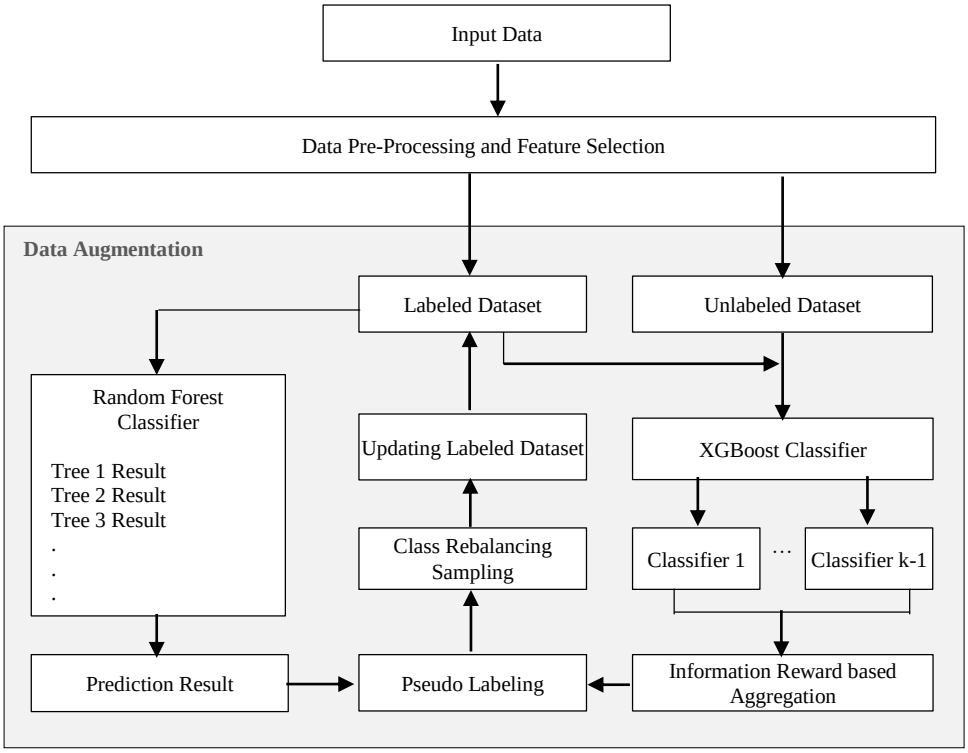


Fig. 2 Detailed working procedure of the proposed model

4. Experimental Analysis

This section discusses the study analysis made for the proposed work by performing various experiments to perform data augmentation intended to reduce the class inequality proportion. The trial outcomes on the suggested model with various datasets are then assessed with the various other models. To perform the investigational study for the proposed semi-supervised ensemble learning with data augmentation for effective analysis of imbalanced data, a cluster of five nodes is employed in which one node is considered the master and the remaining four nodes are acknowledged as slaves.

The hardware specification of the nodes utilized in the experimental study includes Intel Core i3 CPU processor, two cores per processor with 3.00 GHz, 64 GB RAM and 1000 GB Hard Disk with the Open Source Apache Hadoop with Apache Spark software and machine learning library (MLlib) version of 1.2.2, Hadoop distributed File system (HDFS) with a block size of 128 MB. The master node manages the HDFS and controls and directs the slave nodes.

The model employs a MapReduce framework with multiple map phases with a set of nodes and a single reduce phase. However, instead of utilizing the data as it is, a quality improvement framework has been applied that utilizes minimum redundancy maximum relevance to select the relevant attributes and instances [48].

Next, the given input labeled training set is split into several partitions by applying bagging along with the test sample and is then assumed as an input for the nodes in the map phase. The three map nodes apply an XGBoost classifier, and a single node applies an RF classifier. The models are trained through which an information reward is computed, and the prediction is carried out for the test sample.

The classifiers' information reward and the predicted test result are passed to the reduced phase for performing class weight-based aggregation and rebalancing. The result of this proposed model is the prediction of class variables and the data augmentation of the labeled set with minority classes.

Table 1. Dataset employed

Dataset	#Features	#Instances	#Classes
analcat	71	841	4
cjs	10	2796	6

dna	181	3186	3
epsilon	2000	400000	2
gas-grift	129	13910	6
german	24	1000	2
gina	971	3468	2
hill	101	1212	2
madelon	500	2600	2
segment	20	2310	7
steel	27	1941	7
susy	18	5000000	2
synthetic	62	600	7
texture	41	5500	11
vehicle	19	846	4
wdbc	31	569	2

The proposed model has been experimented with 16 datasets extensively utilized for evaluating the efficiency of the semi-supervised ensemble algorithm on big data and employed in this learning model. The particulars, including features and instance counts of various datasets used for the analysis, are given in Table 1. Here, the epsilon dataset is available at LIBSVM [49], and the other 15 datasets are at the UCI data Repository [45, 50].

5. Results and Comparison

The obtained results for the proposed Semi-Supervised Ensemble Learning model (SSEL) are analyzed with various other existing models such as Linear Regression based Stochastic Gradient Descent (LR-SGD), RF, Linear SVM (LSVM), Decision Trees (DT) and Distributed SVM (DSVM) that are implemented in Apache Spark Framework. The details of the proposed model, analyzing the accuracy and execution time for the Epsilon and Susy datasets and the outcomes obtained are shown in Table 2. Figure 3 also shows a graph of the accuracy values acquired and displayed in Table 2.

Table 2. Result comparison on various models

Method	Epsilon Dataset			Susy Dataset		
	Accuracy	Rate of Rising	Time	Accuracy	Rate of Rising	Time
Spark-LR_SGD	49.53	45.16	112	52.43	41.41	220
Spark-RF	62	31.36	182	72.1	19.42	212
Spark-LSVM	86.2	4.56	87	61.3	31.49	235
Spark-DT	66.4	26.48	104	72.6	18.86	219
Spark-DSVM	88.47	2.05	69	83.12	7.11	43
Proposed SSEL	90.32	-	76	89.48	-	68

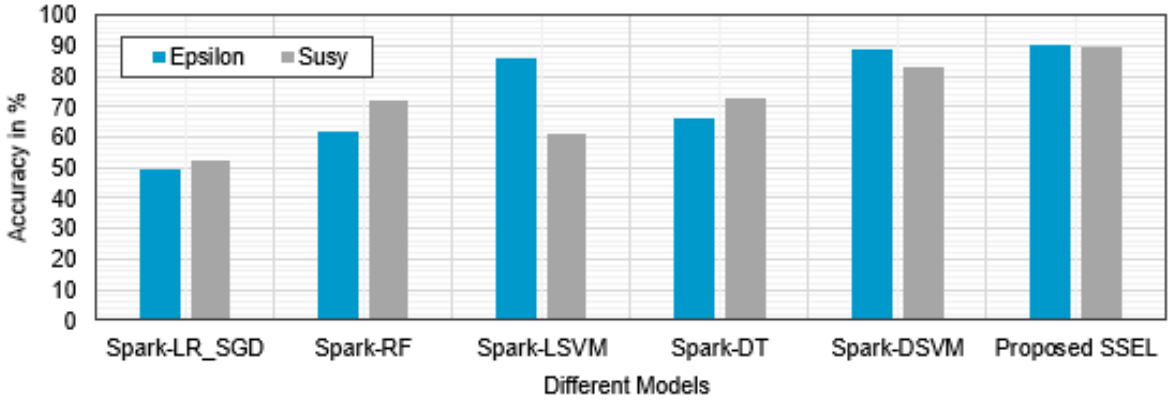


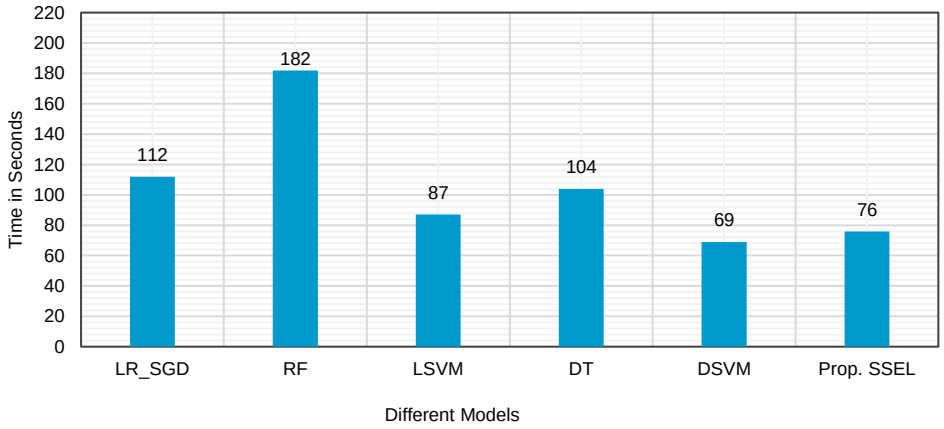
Fig. 3 Accuracy comparison

It is evident from the analysis that the suggested model, the Linear SVM classifier and Distributed SVM offer improved performance with increased accuracy of 86.2% and 88.47% with the Epsilon dataset; however, DSVM has better accuracy of 83.12% for the Susy dataset. On the other hand, the proposed semi-supervised ensemble learning has better accuracy of 90.32% and 89.48% for the Epsilon and Susy datasets. The comparative rate of enhancement in the predictive power of the suggested model with that of other models using the Epsilon dataset ranges from 2% to 45% with an average of 21.9% and that of the Susy dataset ranges from 7% to 41% with an average of 23.65% approximately. The time taken to execute the suggested approach is compared with other models on the Epsilon and Susy datasets and is presented in Figure 4.

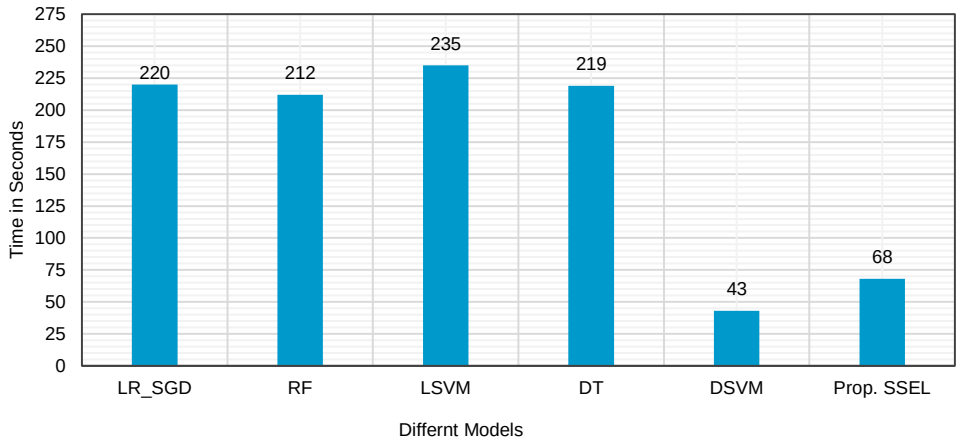
When comparing the execution time of the proposed model, the model exhibits a shorter execution time than several other models, except for the distributed SVM. Though it is found that the proposed model takes more processing

time, compared with the performance concerning the accuracy, the model can be used for sensitive applications where the accuracy is considered most significant and essential.

Similarly, the proposed semi-supervised ensemble learning using XGBoost and RF with class rebalancing sampling is evaluated using 14 other datasets. The findings produced from the suggested model are compared with those of other existing ones that include the Gaussian Mixture Model (GMM) [21], Graph-based SVM (GSVM) [23], Ladder network [25], Co-forest [26], Transductive SVM (TSVM) [22], and Co-Training with Optimal Weight (CTOW) [45]. The accuracy obtained for various classifiers is presented in Table 3.



(a) Epsilon Dataset



(b) Susy Dataset

Fig. 4 Execution time comparison

From the results obtained using 14 datasets, it is clear that GMM and GSVM datasets achieve less accuracy for many datasets, and it takes more time to produce the results for the large datasets, which are not specified in the table. The ladder network model demonstrates high accuracy for texture datasets with many classes, specifically 11. Due to irregular distribution, Co-forest produces the maximum accuracy among other models for the cjs dataset. The TVSM has better results for synthetic datasets with features than instances. With the segment, wdbc, analcat and german datasets, the CTOW model provides better results using XGBoost and TSVM. However, the proposed model utilizing XGBoost and RF with class rebalancing offers improved performance for hill, steel, gina madelon, gas-grift and dna datasets, particularly having more instances. The graph representation for the accuracy comparison for various existing and proposed models is presented in Figure 5. The rate of rise in accuracy for the proposed model compared to the Ladder model is 5.860%, and with Co-forest, the model achieves an increase of about 6.553%. The rate of increase in accuracy is 5.270% for the TSVM classifier, and that of the CTOW model is 1.643%.

Table 3. Accuracy evaluation for different models

Data	GMM	GSVM	Ladder	Co-forest	TSVM	CTOW	Prop. SSEL
cjs	0.293	0.640	0.740	**0.989**	0.654	0.987	0.985
hill	0.488	0.490	0.530	0.492	0.493	0.499	**0.578**
segment	0.694	0.889	0.898	0.907	0.878	**0.925**	0.914
wdbc	0.643	0.940	0.932	0.905	0.949	**0.954**	0.951
steel	0.466	0.627	0.652	0.62	0.673	0.649	**0.723**
analcat	0.206	0.975	0.982	0.876	0.992	**0.993**	0.989
synthetic	0.292	0.908	0.810	0.745	**0.927**	0.920	0.910
vehicle	**0.657**	0.596	0.635	0.631	0.649	0.625	0.613
german	0.614	0.619	0.679	0.712	0.718	**0.716**	0.708
gina	-	-	0.807	0.814	0.835	0.857	**0.866**
madelon	-	-	0.536	0.538	0.518	0.543	**0.614**
texture	-	-	**0.973**	0.877	0.952	0.953	0.948
gas-grift	-	-	0.945	0.927	0.941	0.965	**0.969**
dna	-	-	0.885	0.89	0.894	0.911	**0.921**
Average	0.484	0.743	0.786	0.780	0.791	0.821	0.835

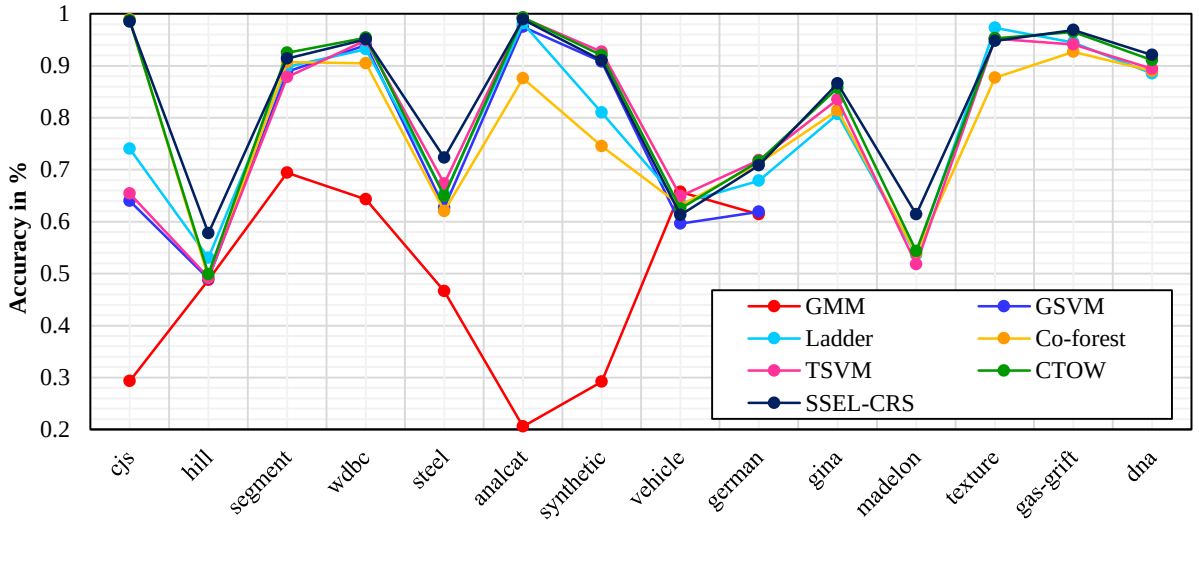


Fig. 5 Accuracy comparison with different models

To evaluate the efficacy of the suggested model and to juxtapose the results with other extant models, the number of datasets exhibiting the most effective accuracy for each model is examined, in which it is found that the suggested model produces the maximum accuracy for 6 datasets. In contrast, CTOW offers improved performance with 4 datasets, and GMM, Co-forest, Ladder, and TSVM models produce improved accuracy for a single dataset. Figure 6 illustrates the distribution of models according to the frequency with which they attain the greatest count, as detailed in Table 3, utilizing various datasets from the study.

An additional analysis was conducted to ascertain the proposed model's performance rate compared to other existing models utilized in the study, focusing on the mean accuracy of the classifiers throughout the 14 datasets, excluding the epsilon and Susy datasets. Here, GMM acquires the least rank with 48.4%, GSVM acquires 6th position with 74.3%, Co-forest in position 5 with 78%, Ladder obtains 4th rank with 78.6%, TSVM attains 3rd rank with 79.1%, CTOW and Proposed model achieve 2nd and 1st position with 82.1% and 83.5% of average accuracy, respectively. The obtained results are presented as a graph in Figure 7.

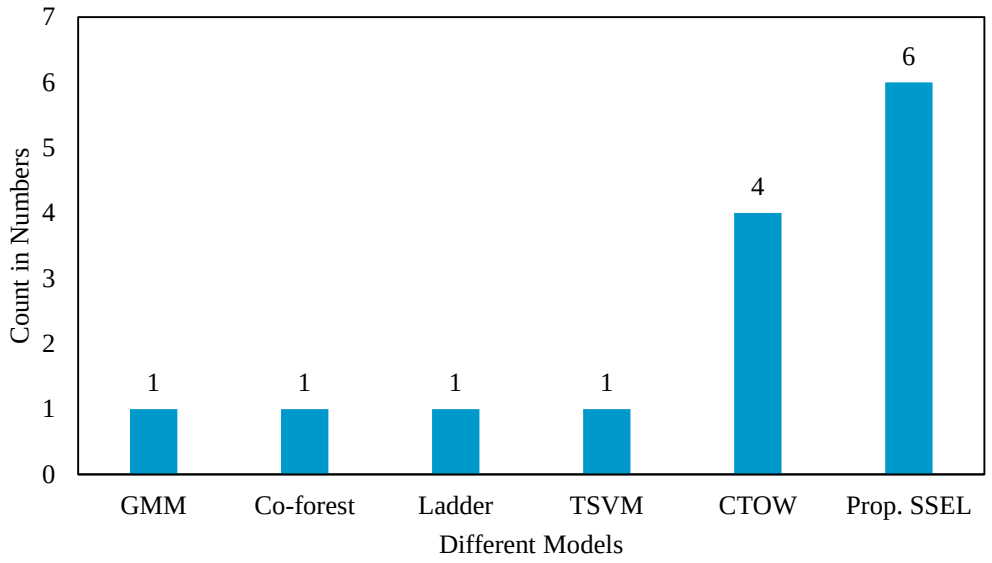


Fig. 6 Distribution of models based on highest accuracy

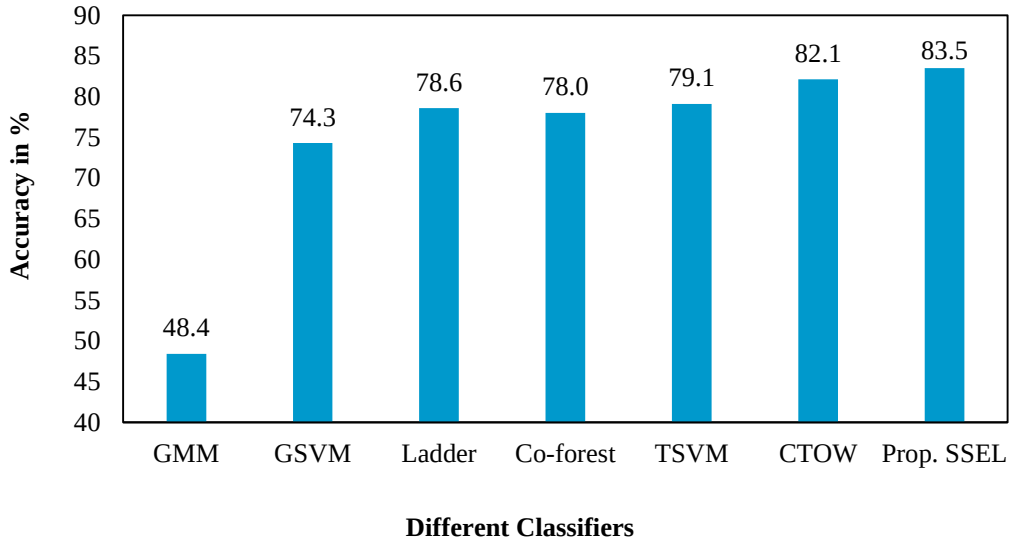


Fig. 7 Average accuracy comparison

From the result analysis performed with various datasets, it is found that the proposed model produces enhanced performance by achieving an improved accuracy of 84.29% for all 16 datasets with an average rate of increase in accuracy of 17.65%.

5. Conclusion

This paper presents a semi-supervised ensemble classification for performing classification on big data with an imbalanced class distribution. The model employs data augmentation by augmenting the instances of the minority class, thereby diminishing the imbalance ratio through class rebalancing sampling. To perform pseudo label, the models apply weight-based aggregation of results from the ensemble classifiers using Bayesian information reward, which are then used to resample the minority class. The experimental test of the suggested model was conducted using 16 datasets, with an average accuracy of 84.3%, much outperforming previous models. With 16 datasets, the suggested approach achieves the highest accuracy for 8 datasets. On average, the proposed model has improved accuracy with an average rate of 17.65% with minimum computational time. However, the proposed method still needs improvement in achieving 100% results. Future work aims to propose a model for achieving 100% results. The model can be extended to use various deep learning models to improve the performance of real-time datasets.

References

[1] Vellingiri Jayagobal, and K.K. Basser, *Data Management and Big Data Analytics: Data Management in Digital Economy*, Research Anthology on Big Data Analytics, Architectures, and Applications, IGI Global, pp. 1614-1633, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Martin Hilbert, and Priscila López, "The World's Technological Capacity to Store, Communicate, and Compute Information," *Science*, vol. 332, pp. 60-65, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mandeep Kaur Saggi, and Sushma Jain, "A Survey towards an Integration of Big Data Analytics to Big Insights for Value-Creation," *Information Processing & Management*, vol. 54, no. 5, pp. 758-790, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Fakhitah Ridzuan, and Wan Mohd Nazmee Wan Zainon, "A Review on Data Quality Dimensions for Big Data," *Procedia Computer Science*, vol. 234, pp. 341-348, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] P.V. Thayyib et al., "State-of-the-Art of Artificial Intelligence and Big Data Analytics Reviews in Five Different Domains: A Bibliometric Summary," *Sustainability*, vol. 15, no. 5, pp. 1-38, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Isaac Triguero et al., "Transforming Big Data into Smart Data: An Insight on the Use of the K-Nearest Neighbors Algorithm to Obtain Quality Data," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 2, pp. 1-24, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Abdulaziz Aldoseri, Khalifa N. Al-Khalifa, and Abdel Magid Hamouda, "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges," *Applied Sciences*, vol. 13, no. 12, pp. 1-33, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jyoti Sharma, "Big Data Management Using Map Reduce Technique," *Academy of Management Annals*, vol. 15, no. 1, pp. 45-51, 2024. [Google Scholar] [Publisher Link]

[9] S.S. Blessy Trencia Lincy, and Suresh Kumar Nagarajan, "A Distributed Support Vector Machine Using Apache Spark for Semi-Supervised Classification with Data Augmentation," *Proceedings of Soft Computing and Signal Processing*, Singapore, pp. 395-405, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Leo L. Duan, James E. Johndrow, and David B. Dunson, "Scaling up Data Augmentation MCMC via Calibration," *Journal of Machine Learning Research*, vol. 19, pp. 1-34, 2018. [Google Scholar] [Publisher Link]

[11] Guillermo Iglesias et al., "Data Augmentation Techniques in Time Series Domain: A Survey and Taxonomy," *Neural Computing and Applications*, vol. 35, pp. 10123-10145, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] S. Sathya Bama, and A. Saravanan, "Efficient Classification Using Average Weighted Pattern Score with Attribute Rank Based Feature Selection," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 7, pp. 29-42, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Aleum Kim, and Sung-Bae Cho, "An Ensemble Semi-Supervised Learning Method for Predicting Defaults in Social Lending," *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 193-199, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kushankur Ghosh et al., "The Class Imbalance Problem in Deep Learning," *Machine Learning*, vol. 113, pp. 4845-4901, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shivani Goswami, and Anil Kumar Singh, "A Literature Survey on Various Aspect of Class Imbalance Problem in Data Mining," *Multimedia Tools and Applications*, vol. 83, pp. 70025-70050, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] N.V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[17] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," *Neural Networks*, vol. 106, pp. 249-259, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] Shuo Wang, Leandro L. Minku, and Xin Yao, "Resampling-Based Ensemble Methods for Online Class Imbalance Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1356-1368, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[19] Saravanan Arumugam, Anandhi Damotharan, and Srividya Marudhachalam, "Class Probability Distribution Based Maximum Entropy Model for Classification of Datasets with Sparse Instances," *Computer Science and Information Systems*, vol. 20, no. 3, pp. 949-976, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Chen Wei et al., "CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 10852-10861, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Xiaojin Zhu, "*Semi-Supervised Learning Literature Survey*," Technical Report, University of Wisconsin-Madison, pp. 1-60, 2005. [Google Scholar] [Publisher Link]

[22] Kristin P. Bennett, and Ayhan Demiriz, "Semi-Supervised Support Vector Machines," *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Cambridge, MA, United States, pp. 368-374, 1998. [Google Scholar] [Publisher Link]

[23] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399-2434, 2006. [Google Scholar] [Publisher Link]

[24] Harri Valpola, *Chapter 8 - From Neural PCA to Deep Unsupervised Learning*, Advances in Independent Component Analysis and Learning Machines, Academic Press, pp. 143-171, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Antti Rasmus et al., "Semi-Supervised Learning with Ladder Networks," *Proceedings of the 29th International Conference on Neural Information Processing Systems*, Montreal Canada, vol. 2, pp. 3546-3554, 2015. [Google Scholar] [Publisher Link]

[26] Ming Li, and Zhi-Hua Zhou, "Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 6, pp. 1088-1098, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[27] Saul Calderon-Ramirez et al., "Correcting Data Imbalance for Semi-Supervised COVID-19 Detection Using X-Ray Chest Images," *Applied Soft Computing*, vol. 111, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Paola Cascante-Bonilla et al., "Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning," *Thirty Fifth Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6912-6920, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Sara del Río et al., "A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules," *International Journal of Computational Intelligence Systems*, vol. 8, pp. 422-437, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[30] Nicholas G. Polson, and Steven L. Scott, "Data Augmentation for Support Vector Machines," *Bayesian Analysis*, vol. 6, no. 1, pp. 1-23, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[31] Panayiota Touloupou et al., "Efficient Model Comparison Techniques for Models Requiring Large Scale Data Augmentation," *Bayesian Analysis*, vol. 13, no. 2, pp. 437-459, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[32] Seth Nabarro et al., "Data Augmentation in Bayesian Neural Networks and the Cold Posterior Effect," *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence PMLR*, UAI, vol. 180, pp. 1434-1444, 2022. [Google Scholar] [Publisher Link]

[33] Giovanna Garcia et al., "An Iterative Heuristic Approach for Channel and Power Allocation in Wireless Networks," *Annals of Telecommunications*, vol. 73, pp. 293-303, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[34] Ranjeet Kumar Singh, "Developing a Big Data Analytics Platform Using Apache Hadoop Ecosystem for Delivering Big Data Services in Libraries," *Digital Library Perspectives*, vol. 40, no. 2, pp. 160-186, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Mingxing Duan et al., "A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2337-2351, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[36] Tianjing Zhao et al., "Fast Parallelized Sampling of Bayesian Regression Models for Whole-Genome Prediction," *Genetics Selection Evolution*, vol. 52, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[37] Tianqi Chen, and Carlos Guestrin, "Xgboost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, pp. 785-794, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[38] Neha Srivastava, and Devendra K. Tayal, "Assessing Gene Stability and Gene Affinity in Microarray Data Classification Using An Extended Relieff Algorithm," *Multimedia Tools and Applications*, vol. 83, pp. 45761-45776, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Yongxiao Qiu, Guanghui Du, and Song Chai, "A Novel Algorithm for Distributed Data Stream Using Big Data Classification Model," *International Journal of Information Technology and Web Engineering*, vol. 15, no. 4, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Huihui Wang et al., "Distributed Classification for Imbalanced Big Data in Distributed Environments," *Wireless Networks*, vol. 30, pp. 3657-3668, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Mikel Elkano et al., "CFM-BD: A Distributed Rule Induction Algorithm for Building Compact Fuzzy Models in Big Data Classification Problems," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 1, pp. 163-177, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Yuzhe Yang, and Zhi Xu, "Rethinking the Value of Labels for Improving Class-Imbalanced Learning," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver BC Canada, pp. 19290-19301, 2020. [Google Scholar] [Publisher Link]

[43] Hyuck Lee, Seungjae Shin, and Heeyoung Kim, "ABC: Auxiliary Balanced Classifier for Class-Imbalanced Semi-Supervised Learning," *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 7082-7094, 2021. [Google Scholar] [Publisher Link]

[44] Fei Sun et al., "An Impartial Semi-Supervised Learning Strategy for Imbalanced Classification on VHR Images," *Sensors*, vol. 20, no. 22, pp. 1-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[45] Zhiguo Wang et al., "Optimally Combining Classifiers for Semi-Supervised Learning," *Arxiv Preprint*, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[46] Lucas R. Hope, and Kevin B. Korb, "Bayesian Information Reward," *Proceedings 15th Australian Joint Conference on AI 2002: Advances in Artificial Intelligence*, Canberra, Australia, pp. 272-283, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[47] A. Saravanan, and Vineetha Venugopal, "Detection and Verification of Cloned Profiles in Online Social Networks Using MapReduce Based Clustering and Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 1, pp. 195-207, 2023. [Google Scholar] [Publisher Link]

[48] A. Saravanan, C. Stanly Felix, and M. Umarani, "Maximum Relevancy and Minimum Redundancy Based Ensemble Feature Selection Model for Effective Classification," *Proceedings of ICACIT Advanced Computing and Intelligent Technologies*, Manipur, India, pp. 131-146, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[49] Chih-Chung Chang, and Chih-Jen Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[50] Dheeru Dua, and Casey A Graff, "*UC Irvine Machine Learning Repository*," University of California, Irvine, School of Information and Computer Sciences, 2017. [Google Scholar] [Publisher Link]