Svedberg Open
DISSEMINATION OF KNOWLEDGE
Research Paper

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Open Access

A Hybrid Big Data Analytics Framework for Business Intelligence Classification Using Whale Optimization Algorithms and Neural Network Models

Dr. Geetha T V[1], R. Naveenkumar[2], Dr.V.Sumathi[3], Ali Bostani[4], Dr.A.Mummoorthy[5], Monisha J[6], Dr .P.Dharmendra Kumar[7]

[1]Assistant Professor, Department of IOT-CSBS/SCSE, SRM Institute of Science and Technology, Ramapuram, Chennai, Email: geethatv.1309@gmail.com, 0000-0002-4809-4996

[2]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India. Email: drnk1983@gmail.com, 0000-0001-9033-9400

[3]Assistant Professor & Head, Department of Computer Technology, Sri Ramakrishna College of Arts & Science (Autonomous), Coimbatore 641006, Email: sumathiviswa@gmail.com, https://orcid.org/my-orcid?orcid=0000-0002-4650-455X

[4]Associate Professor, College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Kuwait, Email: abostani@auk.edu.kw, 0000-0002-7922-9857

[5]Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Email: drmummoorthya@veltech.edu.in, ORCID ID: https://orcid.org/0000-0002-1820-2124

[6]Assistant Professor, Department of Management Studies, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, tamilnadu, India, Email: monishaj@maher.ac.in

[7]Assistant professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India, Email: pdharmendrakumar@kluniversity.iz, dharmendra.phd.au@gmail.com, https://orcid.org/0009-0006-5087-9123

Abstract

Big data plays a significant role in the information industry, and the vast data generation is witnessed by digitization. The process of analysing and handling the issues of velocity, volume, variety, and veracity of the data with the assistance of traditional approaches are ineffective. To overcome the shortcomings of the traditional and other mining approaches, deep learning has initiated to handle the issues of big data. In this research article, the whale optimization algorithm (WOA) is incorporated with an artificial neural network (ANN) for classifying the imbalanced dataset. Thewhale optimization algorithm with the neural network approach (WONNA)is composed of whale optimization, selection of features, pre-processing, and classification. In the optimization phase, whale optimizationremoves the redundancy and irrelevancy of the features where the accuracy of the classification is enriched. The synthetic minority oversampling technique (SMOTE) and synthetic minority oversampling technique with rough set theory with a subset lower approximation (SMOTE-RS B*) is used in the pre-processing and the data is classified in the classification phase. The experimental result shows that the proposed scheme with SMOTE pre-processing has effective performance and promising results when compared to the existing approaches.

Keywords: Big data, neural network, knowledge extraction, optimization, classification, and imbalanced data.

1. Introduction

The extraction of knowledge from a huge database with the available machine learning and classification approach is a complicated process [1]. This recent issue in the information industry is called Big Data that signifies the challenges and difficulties in investigating and analyzingan enormous amount of data. Big has huge consideration in diverse fields' namely electronic commerce, e-health, social networking online, bioinformatics,

and the Internet of Things [2]. These approaches are the key source for the generation of big data and processing these data to retrieve significant insight is a complicated process [3].

The chief characteristic of big data is the vast amount of collected data from diverse sources. Another issue in big data is diverse forms of data namely audio, text, video, hypertext, and image. The structured data can be stored in a two-dimensional format which is simple to store and process. The unstructured form of data is created from mobile devices and the internet that is also in the form of multimedia. The unstructured data is completely different from the conventional format and they need special approaches to process [4].

The velocity of big data necessitates quick and real-time data processing and it is indispensable for many applications. Nevertheless, big data needa robust processing environment and algorithm to acquire significant insight. Traditional mining approaches are insufficient in processing and handling the issues of velocity, volume, variety, and veracity of the data [5]. To rectify the issues, deep learning approaches are introduced to handle the big data problem. In the deep learning method, there are numerous feature selection and learning methods that are effective in handling big data [6].

The learning approaches are enriched by developing numerous approaches by incorporating mining techniques. Artificial Neural Network (ANN) is developed by observing the nervous system and it is determined as a nonparametric mathematical approach [7]. The stochastic and gradient-based approaches are two different types of neural networks. The effectiveness of the algorithm is determined by the count of the iterations needed for finding the different preliminary solution an optimizer necessitates for identifying the global optimum value.

In this paper, the whale optimization is incorporated with an artificial neural network classifying the imbalanced dataset. This proposed approach is composed of three phases namely selection of feature, pre-processing, and classification. In the whale optimization redundancy and irrelevancy of the features are removed and the accuracy of the classification is enriched. The identification of local optimum and parameter optimization is improved by the incorporation of whale optimization in the neural network.

The research article is ordered as follows, existing big data classification approaches and their shortcomings are described in Section 2, the whale optimization with neural network approach is defined in Section 3, experimental results and the relevant discussions are illustrated in section 4 and the proposed WONNA is concluded with a suggestion for future work in Section 5.

2. Related Work

Classification tasks have been extensively performed by using supervised machine learning (SML) methods because they can be trained to learn mapping functions by use of labeled data. The classical models like Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbors (KNN) and Naïve Bayes (NB) have shown good performance in different fields. SVM is especially useful in working with high-dimensional data and building optimum decision boundaries. Nonetheless, as the scale of big data grows exponentially, the traditional SML methods become more complicated in computation and limited in scalability, which necessitates complex optimization policies and computational models [3].

In order to overcome these constraints, quantum machine learning (QML) has been proposed as a new paradigm that increases the efficiency of computations based on quantum principles. Specifically, the Quantum Support Vector Machine (QSVM) is a binary classification model that offers optimization problems with a lower level of computation complexity. Quantum algorithms, including the Harrow-Hassidim-Lloyd (HHL) algorithm, are used to speed up the computationally expensive matrix inversion process of classical SVM. Moreover, non-sparse matrix exponentiation is an essential part of quantum big data analytics, allowing to work with large-scale data and enhance the efficiency of training [3].

Intrusion detection systems (IDS) used in the field of network security are based on the constant monitoring and processing of network traffic data. The unceasing flow of data results in streams of data of high volume, and real-time analysis becomes difficult. It has been processed using machine learning-based methods to derive meaningful patterns to be used in detecting anomalies. Online learning and incremental learning are

some of the techniques used to provide adaptive update in models to achieve effective intrusion detection in dynamic environments. Such methods go a long way in improving the accuracy of detection and addressing the scalability issues of big data [4].

Big data processing and maintenance necessitate strong and scalable models to overcome the limitations of the conventional data processing methods. Models that are based on machine learning offer adaptive learning features that enhance efficiency in data handling. Scalability is however a significant challenge because of the growing size and complexity of datasets. To cope with this problem, distributed computing systems like MapReduce have been developed to use a divide-and-conquer approach. In big data systems, MapReduce breaks down large jobs into small sub-jobs, runs them simultaneously and sums up the output, which greatly enhance scalability and computation in big data systems [5], [6].

Hybrid optimization methods have been widely investigated to improve the performance of classification. By combining Particle Swarm Optimization (PSO) with SVM, the hyperparameters can be tuned effectively and the accuracy of the classification increased by optimization of decision boundaries. In the same vein, Cat Swarm Optimization (CSO) a natural cat behavioral algorithm has been utilized on feature selection and classification tasks, especially on high-dimensional data. Such metaheuristic algorithms are efficient to balance exploration and exploitation, thus leading to better convergence and generalization performance [7], [8].

The Correlative Naive Bayes (CNB) classifier has been developed to be further optimized with the use of advanced hybrid models like the Cuckoo-Grey Wolf Optimization (CGWO) with the aim of improving the efficiency of the classifier. CG-CNB model combines the optimization strategies with the probabilistic classification, in which the posterior probabilities are calculated based on probability index tables. The CGCNB-MRM framework enhances the accuracy of classification by taking into account the correlation of the features and the optimization of the choice of parameters. Also, Scale-Free Particle Swarm Optimization (SF-PSO) has been suggested to be used in feature selection of high-dimensional datasets and it is commonly used together with Multi-Class Support Vector Machine (MC-SVM) to perform successful multi-class classification tasks [9], [10].

Although these improvements have been made, some of the challenges that big data classification approaches continue to encounter are high cost of computation, redundancy of features and scalability. In order to overcome them, the recent studies were aimed at combining deep learning and optimization algorithms, which allowed extracting features automatically and enhancing the performance of the classification process. These hybrid strategies have been shown to be more effective to work with large-scale, high-dimensional data with high-capacity and efficiency, which opens the way to the next generation intelligent data analytics systems.

3. Proposed Phase

The big data classification framework is composed of three phases namely selection of feature, pre-processing, and classification phase. The overall performance of the proposed classification approach is depicted in Figure 1.

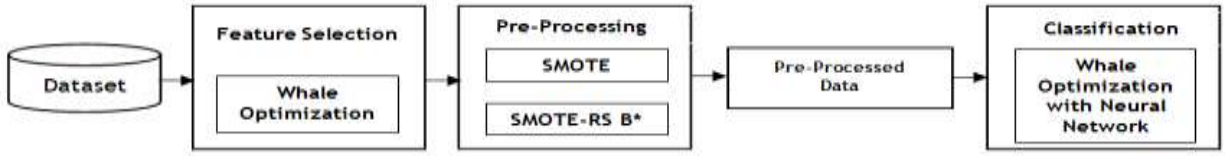

Figure 1: Proposed Classification Approach - WONNA

Whale Optimization Algorithm (WOA)

There are numerous species in the whale namely sei, finback, humpback, and blue where the whales are considered as smartest species and they have spindle cells in their brain. The spindle cells provide assistance in social behavior, emotions, and judging that are similar to humans. Among the varieties of whales, humpback whales have unique hunting behavior called the bubble-net feeding approach. In this approach, circles of bubbles are generated for the hunting process. The hunting is categorized as a double loop and upward spiral

method. The behavior of the bubble net feeding is formulated using the mathematical approach for optimization [16]. The redundancy and irrelevancy of the data are removed by the whale optimization approach. The resultant value is passed to the subsequent phase.

Feature Selection Phase

The classification accuracy has a huge impact with the irrelevant features or redundant whereas the subsets are retrieved to enhance the classification accuracy.Every subset is considered as a whale's position for the selection of relevant features and it holds a certain random number of features that is equal or less than the total count of features in the original dataset. The whale having greater accuracy and minimum count of features is considered an excellent solution.

Pre-processing Phase

The sampling techniques namely SMOTE and SMOTE-RS B* are included for pre-processing. Themain intent of the Synthetic minority oversampling technique (SMOTE) [17]is producing the synthetic samples over the minority class by functioning over the domain attribute instead of domain instance.Every occurrence of the value in the minority class is oversampled by generatingsynthetic instances across the k minority nearest neighbor. The value is selected randomly and five is the default value of k-nearest neighbor.A hybrid under sampling and over sampling approach is by incorporating synthetic minority oversampling technique with rough set theory with a subset lower approximation.SMOTE-RS B* constructs new samples with the unbalanced dataset [18].

Classification Phase

The classification phase elucidates the proposed classification algorithm whale optimization with neural network (WONNA) for identifying the optimal values of biases and weights of WONNA. The needed variables are initialized and the count of the whale is 150 with 20 iterations. Classification accuracy is the same after iteration 20 and increasing the iteration also increases the running time. Hence, the iteration value is assigned as 20 and the count of the whale is elected randomly. In the training process, the metaheuristic approach is applied and the values of bias or weight will influence the accuracy of the classification. The optimal values of bias and weightare identified to attain the highest accuracy. The vector form of whale optimization is given as,

$$\overrightarrow{Vector} = \{\overrightarrow{wt}, \overrightarrow{bs}\} = \{wt_{11}, wt_{12}, \dots\dots, wt_{nn}, k, bs_1, bs_2, \dots\dots, bs_k\}$$

where n denotes the count of input nodes, the connection between the x^{th} node to the y^{th} node is W_{xy} and the bias is signified by bs_y.

The objective function is determined and the estimation metrics used for the evaluation is mean square error [19]. It determines the variation among the proposed classified output and desirable output. Estimation of training instances is equated as,

$$MSE = \sum_{x=1}^{m} (o_x^h d_x^h)^2$$

where m denotes the count of output nodes, $d_x^h$is the optimal output value of the xth input neuron when the training instance h^{th} is applied and o_x^h is considered as the actual output of the x^{th} input neuron when the training instance h^{th} occurs in the input. To make the WONNA classifier as more effective, the performance of the WONNA is estimated under the average value of MSE over all the instances in the training is given as,

$$\overline{MSE} = \sum_{h=1}^{ti} \frac{\sum_{x=1}^{m} (o_x^h d_x^h)^2}{ti}$$

where the training instance is signified as s.

The proposed classification training is formulated with the average MSE and a variable for whale optimization is given as,

$$minimize: f\left(\overrightarrow{vector}\right) = \overline{MSE}$$

The proposed approach is terminated if the process of searching attains maximum iteration that is it retains the same solution over maximum iteration or the best solution is not altered over definite loops.

4. Result and Discussion

The experiment is in the system with 12 GB RAM and Windows operating system with the tool MATLAB R. Performance of the proposed algorithm is compared with the existing classifiers namely random tree, naïve bayes, and decision table. The experiment is carried with big data dataset ECBDL 14 and three more datasets from KEEL, which are all imbalanced. The ECBDL 14 dataset has 631 attributes and 32 million instances whereas the experiment is carried using the 23 attributes and 2,971,986 instances [20] The proposed and existing algorithm is compared without and with the pre-processed dataset. The dataset characteristics are given in Table 1 and the confusion matrix is given in Table 2.

Table 1. Characteristics of Dataset

Dataset Name	Attributes	Examples	IR
Yeast 1	7	459	14.3
Yeast 5	8	1,484	32.73
Yeast 6	8	1,484	41.4
ECBDL 14	23	2,971,986	58.58

Table 2: Confusion Matrix

	Predicted Positive	Predicted Negative
Positive Class	True Positive	False Negative
Negative Class	False Positive	True Negative

4.1. Performance estimation metrics

The general metric used for the evaluation of classification performance accuracy and the value of accuracy is estimated by,

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

For an imbalanced dataset, the estimation of results with accuracy is not accurate and other performance metrics are needed to classify the data over independent classes of negative and positive value. The specificity and sensitivity are accurate in estimating the performance of the classifiers. The sensitivity investigates the capability of identifying the true positive rate and specificity investigates the capability of identifying the true negative rate [21].

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

The mean squared error (MSE) estimates the generation of normal distribution value that estimates the local optima avoidance. The geometric mean of F-measure, accuracy, precision-recall curve, and precision is the area under curve value (AUC) [22]. For an imbalanced dataset, AUC is the best classification indicator and independentamong the classes with distributed instances. The value of AUC is estimated by,

$$AUC = \frac{sensitivity + specificity}{2}$$

For imbalanced problem, Synthetic Minority Over-sampling Technique (SMOTE) and SMOTE-RS B*is applied for pre-processing the datasets. The configuration of SMOTE is with 5 nearest neighbor value and a 50 % oversampling rate with 2 instances per leaf.

4.2. Experimental analysis

In this experiment, the proposed approach is investigated by comparing with the existing classifiers namely random tree, naïve bayes, and decision tablewith the datasets diverse forms of yeast and ECBDL 14. The comparison is made for the pre-processed and un-processed data where the comparison is illustrated in the following Table and Figure.

Table 3: Comparison of AUC value without processing the dataset

Datasets	Classifiers			
	Random Tree	Naïve Bayes	Decision Table	WONNA
Yeast 1	0.561	0.803	0.531	0.932
Yeast 5	0.641	0.971	0.811	0.991
Yeast 6	0.731	0.928	0.562	0.941
ECBDL 14	0.581	0.681	0.557	0.753

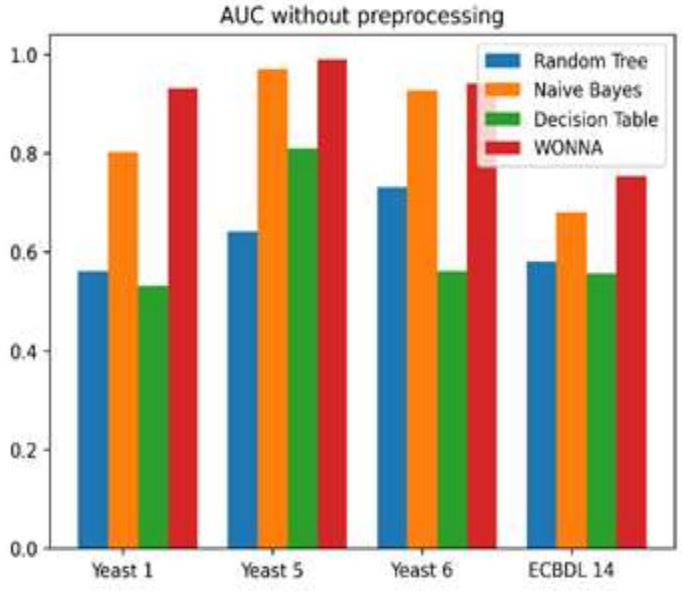

Figure 2: Comparison of AUC value without processing the dataset

In Table 3 and Figure 2, a comparison of AUC without applying any pre-processing approaches, and the proposed approach acquired the highest AUC value. It shows that the proposed algorithm has the best result.

Table 4: Comparison of MSE value without processing the dataset

Datasets	Classifiers			
	Random Tree	Naïve Bayes	Decision Table	WONNA
Yeast 1	0.169	0.207	0.167	0.0394
Yeast 5	0.238	0.076	0.034	0.0059
Yeast 6	0.0813	0.023	0.078	0.0056
ECBDL 14	0.0284	0.196	0.172	0.0991

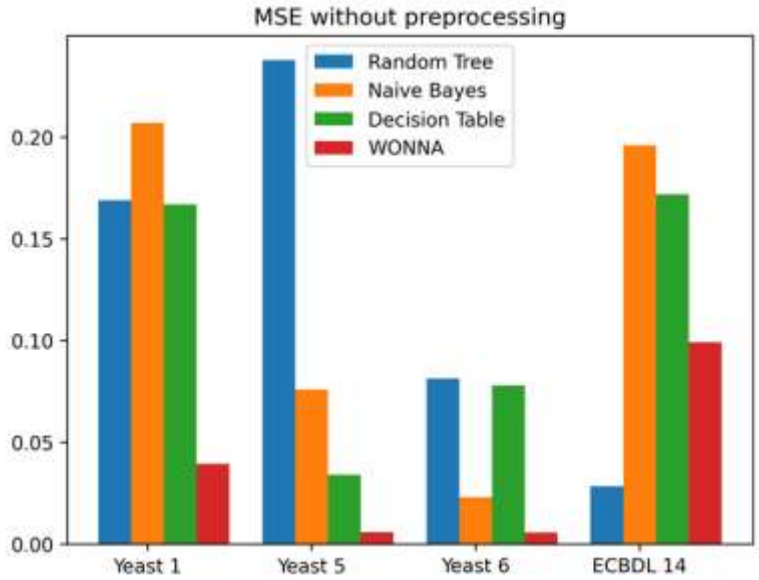

Figure 3: Comparison of MSE value without processing the dataset

In Table 4 and Figure 3, a comparison of MSE without applying any pre-processing approaches, and the proposed approach acquired the lowest MSE value. It shows that the proposed algorithm has the best result. The lowest MSE signifies the values are closely dispersed and are effective.

Table 5: Comparison of AUC value with SMOTE pre-processing the dataset

Datasets	Classifiers							
	Random Tree		Naïve Bayes		Decision Table		WONNA	
	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*
Yeast 1	0.952	0.941	0.922	0.812	0.871	0.771	0.991	0.891
Yeast 5	0.765	0.665	0.785	0.725	0.821	0.801	0.912	0.802
Yeast 6	0.862	0.761	0.872	0.732	0.918	0.888	0.931	0.901
ECBDL 14	0.877	0.717	0.894	0.714	0.932	0.892	0.963	0.891

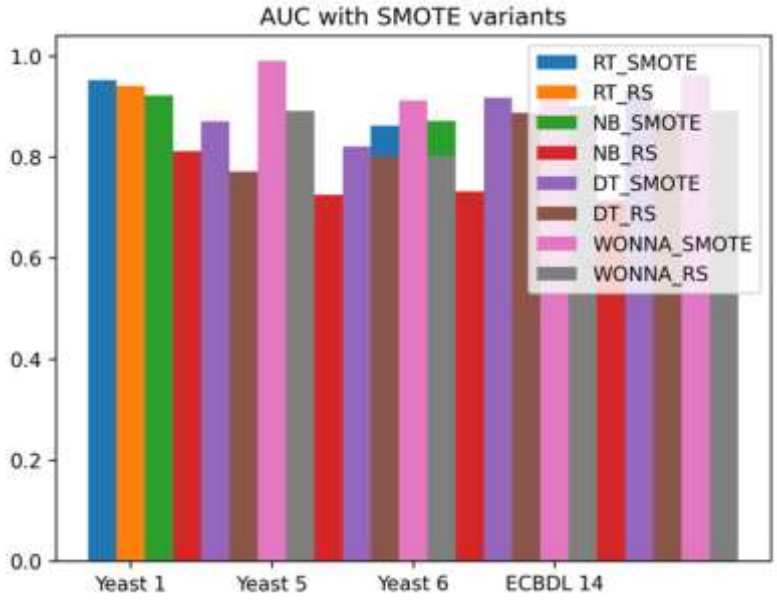

Figure 4. Comparison of AUC value with SMOTE and SMOTE-RS B* pre-processing the dataset

In Table 5 and Figure 4, a comparison of AUC by applying the pre-processing approaches namely SMOTE and SMOTE-RS B*. The proposed approach acquired the highest AUC value and it shows that the proposed algorithm has the best result. From the observation of results, the SMOTE pre-processing is effective when compared to the SMOTE-RS B*.

Table 6: Comparison of MSE value with SMOTE pre-processing the dataset

Datasets	Classifiers							
	Random Tree		Naïve Bayes		Decision Table		WONNA	
	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*	SMOTE	SMOTE-RS B*
Yeast 1	0.0369	0.0261	0.0207	0.0201	0.0294	0.0191	0.017	0.0257
Yeast 5	0.0231	0.0193	0.0176	0.0146	0.0019	0.0011	0.021	0.019
Yeast 6	0.0913	0.0811	0.087	0.047	0.064	0.054	0.051	0.049
ECBDL 14	0.0284	0.0212	0.0146	0.0116	0.0191	0.0171	0.0152	0.0112

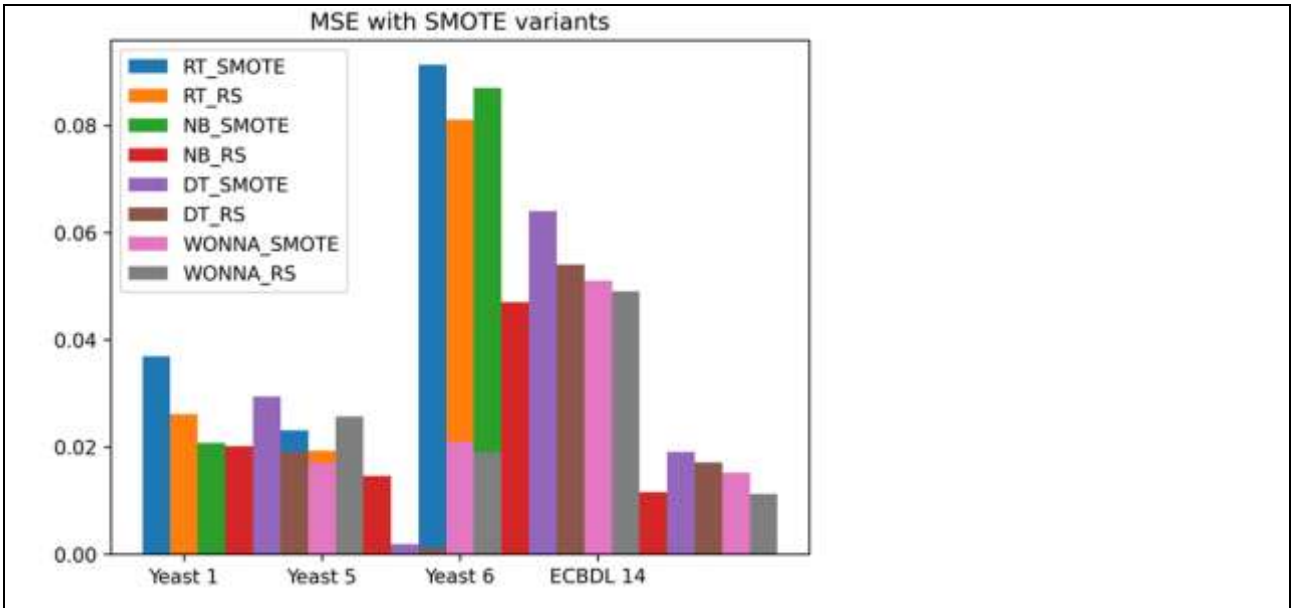

Figure 5: Comparison of MSE value with SMOTE and SMOTE-RS B* pre-processing the dataset

In Table 6 and Figure 5, a comparison of MSE by applying the pre-processing approaches namely SMOTE and SMOTE-RS B*. The proposed approach acquired a minimum MSE value and it shows that the proposed algorithm has the best result. From the observation of results, the SMOTE pre-processing is effective when compared to the SMOTE-RS B*.

Table 7: Comparison of running time in minutes with the diverse pre-processing approach

Dataset	SMOTE Pre-Processing	SMOTE-RS B*
Yeast 1	3.3	4.1
Yeast 5	5.16	7.72
Yeast 6	6.61	9.67
ECBDL 14	491	532

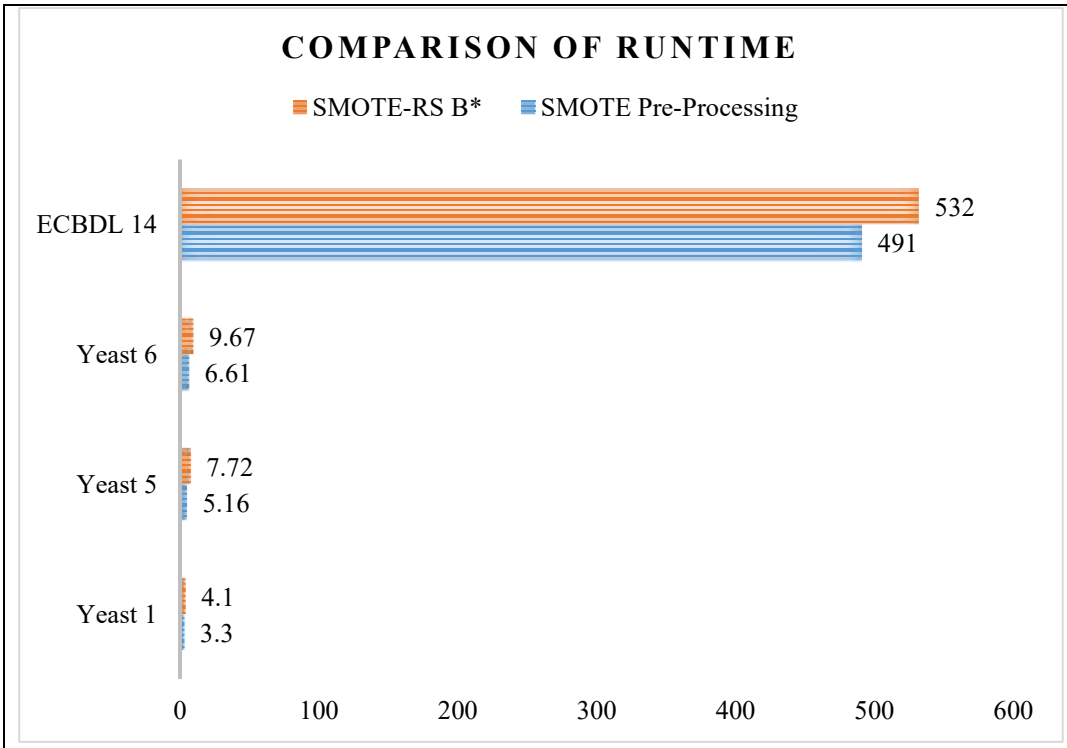

Figure 6: Comparison of running time in minutes with the diverse pre-processing approach

In Table 7 and Figure 6, the execution time for the pre-processing approach SMOTE and SMOTE-RS B* is compared. The SMOTE is effective by acquiring the results in minimum execution time.

5. Conclusion

In this research article, a whale optimization-based artificial neural network approach is developed for an imbalanced dataset. Initially,the irrelevant features are removed by the whale optimization through this best classification accomplished. The synthetic minority oversampling technique (SMOTE) and synthetic minority oversampling technique with rough set theory with a subset lower approximation (SMOTE-RS B*) applied in the pre-processing stage and the data is classified from the imbalanced dataset. The proposed approach is implemented three datasets from KEEL and one big data. The performance is evaluated using the parameters AUC and MSE. The experiment is carried for both imbalanced and pre-processed data. The experimental result shows that the proposed scheme with SMOTE pre-processing has effective performance and promising results when compared to the existing approaches. The proposed approach in the big data has a slightly higher run time and in the future, it is rectified by the algorithm on Hadoop or Spark framework.

Reference

1. Bennin, K. E., Keung, J., Phannachitta, P., Monden, A., & Mensah, S. (2017). Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. *IEEE Transactions on Software Engineering*, *44*(6), 534-550.
2. Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, *106*, 15-29.
3. Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical review letters*, *113*(13), 130503.
4. Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), 70-73.
5. Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst*, *36*, 1-12.

6. Fernández, A., del Río, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, *3*(2), 105-120.
7. Demidova, L., Nikulchev, E., & Sokolova, Y. (2016). Big data classification using the SVM classifiers with the modified particle swarm optimization and the SVM ensembles. *International Journal of Advanced Computer Science and Applications*, *7*(5), 294-312.
8. Lin, K. C., Zhang, K. Y., Huang, Y. H., Hung, J. C., & Yen, N. (2016). Feature selection based on an improved cat swarm optimization algorithm for big data classification. *The Journal of Supercomputing*, *72*(8), 3210-3221.
9. Banchhor, C., & Srinivasu, N. (2020). Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification. *Data & Knowledge Engineering*, *127*, 101788.
10. Gupta, S. L., Baghel, A. S., & Iqbal, A. (2019). Big data classification using scale-free binary particle swarm optimization. In *Harmony Search and Nature Inspired Optimization Algorithms* (pp. 1177-1187). Springer, Singapore.
11. Latha B. (2025). Intelligent Control Strategies for AC–DC–AC Power Conversion in Electric Drive and Sustainable Power Systems. *National Journal of Electrical Machines & Power Conversion*, 31-39.
12. Haitham M. Snousi. (2025). FPGA-Accelerated Clinical Workflow Orchestration Using Causal Graph Pipelines for Distributed Hospitals. *Journal of Reconfigurable Hardware Architectures and Embedded Systems*, *2*(3), 10–18.
13. Shaik Sadulla. (2025). Foundation Model–Powered Medical Dialogue Agents with Causality-Aware Workflow Reasoning. *Journal of Intelligent Assistive Communication Technologies*, *2*(1), 49-56.
14. T M Sathish Kumar, & Shaik Sadulla. (2025). Fractional-Order Mathematical Models for Vibration Analysis in Smart Structural Systems. *Journal of Applied Mathematical Models in Engineering*, *1*(4), 1–8.
15. Naren Swamy Jamithireddy. (2026). Balancing Cultural Heritage Preservation and Sustainable Tourism Development: A Data-Driven Framework for Smart Destination Management in Emerging Economies. *Journal of Tourism, Culture, and Management Studies*, *3*(1), 22–29.