Adaptive Feature Selection Of Unbalanced Data For Skiing Teaching

Tao Feng*

Physical Education Teaching and Research Department, Harbin Finance University, Harbin 150030 China

Corresponding author. E-mail: 910675024@qq.com

Received: May 14, 2025; Accepted: Jun. 16, 2025

In skiing teaching, the data may show an unbalanced distribution. For example, the sample size of some common movements (such as straight downhill) may be much larger than that of some difficult movements (such as aerial spins). If the features are not selected, the model may overly rely on the features of common actions and ignore the features of difficult actions. Through the feature selection of unbalanced data, features that have significant influences on different actions (especially a few types of actions) can be screened out, thereby enhancing the recognition ability and generalization ability for various actions. The high-dimensional characteristics of the skiing dataset will reduce the classification effect of unbalanced learning. Aiming at the classification problem of high-dimensional unbalanced data, this paper proposes an adaptive feature selection method. This new algorithm combines embedded and wrapped feature selection methods and is capable of adaptively selecting the optimal features to form the feature space. Finally, the experimental results on the public imbalanced dataset show that the proposed algorithm effectively improves the classification performance of imbalanced data.

Keywords: skiing teaching; adaptive feature selection; unbalanced data; high-dimensional characteristics

© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202603_29(3).0006

1. Introduction

In skiing teaching, time and energy are limited. Through feature selection, coaches can focus more energy on the features that have the greatest impact on trainees. For instance, if a certain feature (such as the bending Angle of the knee) is crucial for the completion of difficult movements, the coach can focus on guiding the trainee to adjust this feature instead of applying force evenly across all the details. This not only improves the teaching efficiency, but also reduces the learning burden of the trainees [1]. When the model can identify the movement problems of trainees more accurately, trainees can obtain more timely and targeted feedback. This kind of personalized guidance can enhance the learning experience of trainees and boost their interest and confidence in learning. Furthermore, by optimizing feature selection, coaches can avoid providing unnecessary guidance to trainees, reduce their frustration, and thereby better motivate them to continue learning [2, 3].

Unbalanced data is widely present in practical applications. In the classification task of imbalanced data, the number of samples between classes varies greatly [4]. The goal of traditional learning algorithms is to minimize classification errors and maximize overall accuracy, and usually a relatively balanced category distribution needs to be assumed. Therefore, training models on unbalanced data usually tends to favor the majority of classes while ignoring a few, which can easily lead to deviations in the overall classification performance. The problem of unbalanced data classification is usually solved from the data level and the algorithm level. Data-level methods do not rely on classifier models in the data preprocessing process and have stronger general capabilities. They mainly use strategies such as oversampling, under-sampling, and mixed sam-

pling [5]. Typically, oversampling methods tend to exaggerate the possibility of over-fitting, while under-sampling methods can lead to the loss of some key data. In contrast, hybrid sampling can effectively reduce the impact of a single oversampling or under-sampling technique. Therefore, this paper adopts a mixed sampling strategy to balance the number of samples of various types.

Neighborhood rough sets are widely used to handle symbolic and continuous numerical data for feature selection by granulating the data in the neighborhood [6]. However, it cannot describe the fuzzy similarity among samples in a fuzzy background. To address this limitation, fuzzy neighborhood rough sets are used to handle these fuzzy and uncertain data and noise. Furthermore, it can construct robust distance functions, describe sample decisions using the granularity of fuzzy information, and reduce the error rate of classifying complex data [7]. For example, Liu et al. [8] used neighborhood rough sets to perform feature selection on multi-label data. However, the neighborhood radius needed to be manually selected, consuming a large amount of computational costs. Sang et al. [9] constructed an adaptive fuzzy neighborhood strategy for feature selection. However, the fuzzy similarity relationship between samples only considered the linear relationship in the feature space and ignored the nonlinear relationship, which could lead to the reduction of classification ability. These traditional neighborhood rough set models cannot exhibit good classification performance when directly dealing with imbalanced data.

From the perspective of clustering, the study of feature selection can enhance the classification ability by deleting irrelevant or redundant features. Niño-Adan et al. [10] considered the influence of the features on their nearest neighbor features, clustered the features based on the weighted K-nearest neighbor density, and selected the feature with the maximum redundancy from each feature cluster to form the feature subset. Sun et al. [11] proposed a feature selection algorithm based on similarity-based adaptive weighted K-nearest neighbor feature clustering for imbalanced data. In each feature cluster, the features that were strongly related to the labels and highly redundant with the features in the same cluster were selected to form the final feature subset. However, these two algorithms do not consider the combined performance of the features selected from different clusters. Chormunge and Jena [12] proposed a feature reduction algorithm based on correlation, using K-means clustering to select important features. Dhabliya et al. [13] introduced cosine similarity into K-means clustering to select important features. However, the classification results of both are affected by the number K value in the k -means clusters.

To sum up, the existing feature selection methods for solving high-dimensional unbalance problems still have problems such as the tendency to mistakenly delete key features when constructing sub-spaces, the insufficiently adaptive process of forming feature spaces, and difficulty in being applicable to features of different dimensions. Therefore, based on the Bagging ensemble framework, this paper proposes a feature selection algorithm combining embedded and wrapped methods, and obtains the adaptive optimal feature space by setting the feature extraction rate. In the experimental part, we conduct comparative analysis with different feature selection methods, and explore the optimal sampling methods and the optimal feature extraction rates of high-dimensional datasets with different dimensions and different imbalance ratios, providing a referenceable value range for related studies.

2. Materials and methods

2.1. Proposed algorithm process

The prerequisite for the effectiveness of ensemble learning is to construct diverse and accurate base classifiers, while random forest ensures the diversity of the generated base classifiers by introducing dual randomness in samples and features. Meanwhile, during the construction process of the decision tree, each node of the tree selects important features from numerous random features based on certain principles. This process is actually an explicit feature selection process. Random forest is an ensemble of decision trees and thus correspondingly inherits the ability of decision trees to select important features [14, 15].

However, when constructing the decision tree in the random forest, the best splitting point is selected in the random subspace. Therefore, it is very likely to incorporate redundant or noisy features, making it difficult to guarantee the accuracy of the base classifier. From this perspective, the accuracy of the base classifier can be improved by controlling the feature space of the constructed decision tree. If the importance ranking of each feature is known in advance, then the least important feature can be removed and trained with a decision tree. The recursive feature elimination algorithm based on the wrapper method can meet this requirement. If one wants to improve accuracy while ensuring randomness, it is only necessary to introduce a feature selection process before constructing the decision tree.

Fig. 1 shows the flowchart of the proposed feature selection algorithm for unbalanced data. The original data is divided into the imbalanced training set X^{train} and the test set X^{test} . For the training set, the sampling algorithm

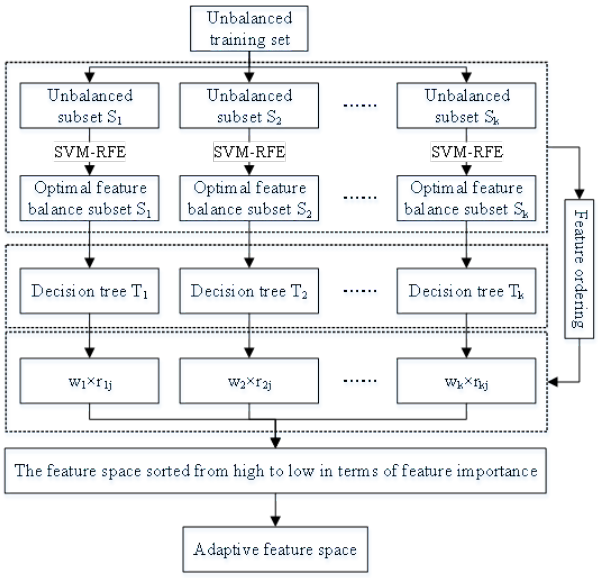

Fig. 1. Proposed feature selection method for unbalanced data

is used to generate multiple balanced training subsets $S = \{S_1, S_2, \cdots, S_T\}$. The advantage of this processing is that multiple different sample subsets are constructed, ensuring the randomness of the samples and containing more information than on a single sample subset. In addition, different sampling algorithms can also be combined. The following text will conduct combined experiments by combining different sampling algorithms to observe the classification effect of the algorithm under different sampling methods.

For each balanced subset, SVM-RFE is adopted for training. SVM-RFE is a feature selection method that combines the SVM algorithm with the recursive feature elimination method. Thus, a set of features ranked by importance is obtained. The most important part of the features is selected to train the decision tree. The training effect of the decision tree on the test set is taken as the weight of the feature ranking. The weighted sum can obtain the total feature importance ranking.

2.2. Feature importance and weight measurement

The measurement of feature importance in the proposed algorithm is calculated based on SVM-RFE. In the SVM-RFE algorithm, features can be sorted based on the size of w^2, and the feature with the smallest w^2 can be deleted each time, thereby obtaining the corresponding feature sorting. It can be assumed that the ranking of feature j after the balanced subset of group i is trained by SVM-RFE is r_j^i.

After going through the above process, each balance sub-training set will generate a feature ranking of each feature. When the data unbalance ratio is relatively high, the differences among the generated balanced subsets will also increase accordingly [16].

Therefore, the classification effects of SVM established on different balanced subsets will also be different. Thus, the obtained feature importance rankings should be assigned with different weights. From this, it can be known that if the training set is classified on the feature space selected by SVM-RFE, the classification effect is better, the feature space selected by this model will be better, that is to say, it should be assigned with a higher weight. In order to verify the feature selection effect of each balanced subset SVM-RFE, the features in the top 50% of the ranking are uniformly selected for the next level of training.

Considering the execution efficiency of Bagging integration, in this paper, the decision tree is used as the base classifier to measure the feature selection effect of SVM-RFE. Decision tree selects the optimal partitioning attributes based on the "purity" of nodes. Several common indicators for measuring purity include information gain, Gini coefficient, etc [17–19]. This paper adopts the Gini coefficient as the criterion for attribute division of decision trees. This is because, compared with the Gini coefficient, information entropy is more sensitive to impurity. When information entropy is used as an index, the growth of decision trees will be more refined. Therefore, for high-dimensional or highly noisy data, information entropy is prone to overfitting, while the Gini coefficient often performs relatively well in such cases. F1-score is a very good indicator for evaluating the classification effect of imbalanced datasets. In this paper, the decision tree T_i is used to train the balanced subset after feature selection, and the value w_i of F1-score is taken as the weight of SVM-RFE. After weighted summation of the feature importance rankings of all balanced subsets, the final feature importance ranking of each feature can be expressed as.

$$FIM_j = \sum_{i=1}^{k} w_i \times r_j^i \tag{1}$$

$$w_i = F1 - \text{score}\left(T_i, X^{test}\right) \tag{2}$$

Here, w_i represents the classification effect of the $i-th$ decision tree T_i on the test set. r_j^i represents the feature ranking of feature j after the $i-th$ balanced subset undergoes SVM-RFE. By setting the feature extraction rate γ, the proposed algorithm can adaptively select the features with the highest importance ranking among the top $L \times \gamma$ features to form the feature space. The feature extraction rate γ is a very important parameter in the method proposed

Table 1. Data sets description

Number	Data set	Sample	Feature	Category	P/%	N/%
1	Arrhythmia	452	280	13	9.73	90.27
2	Glass	210	10	6	32.71	67.29
3	Heart	270	14	2	44.44	55.56
4	Ionosphere	351	35	2	35.9	64.1
5	Pima	768	11	2	34.9	65.1
6	Tic-Tac-Toe	958	10	2	34.66	65.34
7	vehicle3	846	19	4	25.06	74.94
8	Wdbc	569	31	2	37.26	62.74
9	Wpbc	198	31	2	23.74	76.26
10	yeast1	1484	9	10	28.91	71.09
11	Segmentation	2310	20	7	14.25	85.75
12	DLBCL	77	5470	2	24.68	75.32
13	Lung	203	12601	2	8.37	91.36
14	Breast	84	9126	5	11.9	88.1
15	SRBCT	63	2309	4	13.25	86.75

in this paper. The smaller the γ denotes the greater differences among each feature subspace. The larger γ denotes that the it covers more information. Algorithm 1 shows the complete steps of the proposed feature selection.

Algorithm 1. Adaptive feature selection of unbalanced data

Input: training set X^{train}, random tree number k, data dimension L , feature extraction rate γ.

(1) For $i \in [1, \cdots, k]$, the balanced subset S_i is generated using RUS, and the SVM classifier is trained with S_i to obtain the weight coefficient w_i^2 of each feature. Find the feature with the smallest score in the current feature set, delete this feature until all features are removed to obtain the feature ranking r_j^i. Select the top $L \times 50\%$ features to form the optimal feature subset. Train the decision tree T_i on this optimal feature subset, and the value w_i of F1-score of the random tree T_i on the test set X^{test}. End For
(2) Calculate the final feature importance $F = \sum_{i=1}^{k} w_i r_j^i$.
(3) Sort by the importance of features from high to low.
(4) The first $d = L \times \gamma$ features are extracted to form the feature space.

Output: Adaptive feature space.

3. Results and discussions

The experiment is configured with Windows 10, Intel® Core i7 CPU @ 3.20 GHz and 16 GB of RAM, and is programmed and implemented on MATLAB R2020b. In order to verify the effectiveness of the proposed algorithm, selecting the 15 unbalance experiment data sets from UCI database (https://archive.ics.uci.edu/ml) [20]. The specific information is shown in Table 1, where P and N represent the proportions of minority class samples and majority class samples to the total samples.

Table 2. The optimal γ values of the proposed algorithm on 11 UCI datasets under 3 classifiers

Data set	J48	Random forest	SVM
Arrhythmia	0.1	0.1	0.1
Glass	0.3	0.3	0.6
Heart	0.7	1.0	1.0
Ionosphere	0.1	0.6	0.6
Pima	0.6	0.6	0.6
Tic-Tac-Toe	0.7	0.7	0.7
vehicle3	0.5	0.4	0.1
Wdbc	0.4	1.0	0.3
Wpbc	0.5	0.3	1.0
yeast1	0.4	0.4	0.4
Segmentation	0.8	0.3	0.8

In order to verify the classification performance of the proposed algorithm, the proposed algorithm is compared with other seven feature selection algorithms including the original data processing algorithm (ODP) [21], and neighborhood rough set-based heterogeneous featuresubset selection algorithm (FARNeM) [22], weighted attribute reduction algorithm (WARA) [23], CfsSubsetEval in the Weka workbench (CfsSubsetEval) [24], RSFSAID algorithm [25], symmetrical uncertainty and harmony search-based feature selection algorithm (SYMON) [26], and FRSA algorithm [27]. Meanwhile, 5-fold cross-validation is adopted to analyze the performance of these comparison algorithms. We select the top 11 UCI datasets from Table 1, and use the AUC under the three classifiers (J48, Random Forest and SVM) as the evaluation indicators for classification

Table 3. The AUC values of 8 algorithms under J48 classifier on 11 unbalanced datasets

Data set	ODP	FARNeM	WARA	CfsSubsetEval	RSFSAID	SYMON	FRSA	Proposed
Arrhythmia	0.6286	0.8111	0.7306	0.7265	0.7895	0.6869	0.8124	0.8328
Glass	0.7756	0.8330	0.7979	0.8020	0.8041	0.8222	0.9102	0.8994
Heart	0.8012	0.8202	0.8154	0.8049	0.8278	0.8216	0.7966	0.8014
Ionosphere	0.8942	0.9244	0.8906	0.8784	0.9203	0.8939	0.9289	0.9399
Pima	0.7293	0.7769	0.7268	0.7660	0.7679	0.7559	0.6221	0.8004
Tic-Tac-Toe	0.8834	0.4977	0.8834	0.7305	0.8774	0.8840	0.5731	0.9536
vehicle3	0.7055	0.7449	0.7105	0.6449	0.7587	0.7525	0.9273	0.9579
Wdbc	0.9292	0.9541	0.9357	0.9373	0.9444	0.9391	0.9047	0.9771
Wpbc	0.5769	0.7059	0.5769	0.5583	0.6876	0.6524	0.5793	0.8671
yeast1	0.7001	0.7018	0.7111	0.7394	0.7324	0.7304	0.9991	0.8499
Segmentati on	0.9801	0.9853	0.9859	0.9739	0.9924	0.9876	0.9982	0.9774
Average	0.7822	0.7959	0.7968	0.7784	0.8275	0.8115	0.8229	0.8961

Table 4. The AUC values of 8 algorithms under Random forest classifier on 11 unbalanced datasets

Data set	ODP	FARNeM	WARA	CfsSubsetEval	RSFSAID	SYMON	FRSA	Proposed
Arrhythmia	0.9218	0.9149	0.9408	0.9174	0.9393	0.9267	0.8492	0.9555
Glass	0.9342	0.9391	0.9316	0.9311	0.9377	0.9376	0.9642	0.9703
Heart	0.8855	0.8937	0.8879	0.8779	0.8888	0.8808	0.8865	0.8575
Ionosphere	0.9805	0.9815	0.9797	0.9718	0.9812	0.9758	0.9827	0.9854
Pima	0.8195	0.8223	0.8163	0.7923	0.8244	0.8153	0.6221	0.8317
Tic-Tac-Toe	0.9947	0.4983	0.9941	0.8193	0.9662	0.9586	0.5991	0.9801
vehicle3	0.8669	0.8706	0.8675	0.7170	0.8706	0.8685	0.9752	0.9977
Wdbc	0.9896	0.9917	0.9908	0.9887	0.9923	0.9920	0.9942	1.0000
Wpbc	0.6776	0.7532	0.6820	0.6280	0.7136	0.7239	0.7088	0.8671
yeast1	0.7939	0.7963	0.7953	0.7726	0.7954	0.7918	0.9311	0.9760
Segmentati	0.9998	0.9998	0.9997	0.9974	0.9999	0.9999	0.9999	0.9996
Average	0.8967	0.8601	0.8987	0.8558	0.9009	0.8974	0.8648	0.9474

Table 5. The AUC values of 8 algorithms under SVM classifier on 11 unbalanced datasets

Data set	ODP	FARNeM	WARA	CfsSubsetEval	RSFSAID	SYMON	FRSA	Proposed
Arrhythmia	0.5001	0.5001	0.5001	0.5216	0.7087	0.5001	0.5022	0.8818
Glass	0.6370	0.6690	0.6370	0.6443	0.6949	0.5001	0.9221	0.9120
Heart	0.5259	0.5526	0.5259	0.7468	0.8176	0.8168	0.6084	0.8887
Ionosphere	0.9004	0.9225	0.9004	0.9101	0.9282	0.8625	0.9317	0.9513
Pima	0.5001	0.6135	0.5001	0.4851	0.5906	0.7050	0.5272	0.7719
Tic-Tac-Toe	0.8040	0.5018	0.8040	0.6716	0.7452	0.7239	0.5141	0.7195
vehicle3	0.5001	0.5904	0.5001	0.5048	0.6085	0.5001	0.6093	0.9868
Wdbc	0.5001	0.9405	0.7484	0.5072	0.9466	0.9454	0.8614	1.0000
Wpbc	0.5001	0.5786	0.5002	0.5022	0.5667	0.5001	0.5813	0.7574
yeast1	0.5163	0.5361	0.5198	0.5405	0.5519	0.5619	0.7242	0.9063
Segmentati	0.6597	0.8849	0.9707	0.7304	0.9895	0.9791	0.8357	0.9881
Average	0.5949	0.6626	0.6461	0.6150	0.7408	0.6905	0.6925	0.8876

performance. The larger AUC value denotes the better classification performance. The variation range of the feature extraction rate γ used in the proposed algorithm is $[0.1, 1]$. The optimal values corresponding to the feature extraction rate γ of the proposed algorithm under three different classifiers are shown in Table 2. Tables 3 to 5 present the AUC values with 8 algorithms under 3 classifiers on 11 datasets.

According to the above results, it can be known that under the J48 classifier, the AUC values of the proposed algorithm on the seven datasets (Arrhythmia, Ionosphere, Pima, Tic-Tac-Toe, vehicles3, Wdbe and Wpbe) are all the largest. On the Glass and yeast1, the AUC value of the proposed algorithm is slightly lower than that of the FRSA algorithm, but higher than that of the other six comparison algorithms. On the Heart dataset, the AUC value of the proposed algorithm is slightly lower than that of the

RSFASID, FARNeM, WARA and SYMON algorithms, but higher than that of the ODP, CfsSubetEval and FRSA algorithms. Under the Random Forest classifier, although the AUC value of the proposed algorithm is lower than that of the comparison algorithms on the Heart dataset, on the other 9 datasets, the AUC values of the proposed algorithm are all higher than those of the other 7 comparison algorithms. Under the SVM classifier and on the Glass dataset, the AUC value of the proposed algorithm is slightly lower than that of the FRSA algorithm, but higher than that of the other six comparison algorithms. On the Tic-Tac-Toe dataset, the AUC value of the proposed algorithm is higher than that of the three algorithms, namely FARNeM, CfsSubetEval and FRSA. On the Segmentation dataset, the AUC value of the proposed algorithm is slightly inferior to that of the RSFSAID algorithm. On the other eight datasets, the AUC values of the proposed algorithm are all higher than those of the other seven comparison algorithms. Under the three classifiers, the proposed algorithm performs poorly on the Segmentation dataset. The reason for this is that the proposed algorithm selects redundant features.

4. Conclusions

In imbalanced data, the model may overfit the majority of class actions and underlearn the minority of class actions. Through feature selection, the focus of the model on different types of actions can be balanced, reducing the bias of the model. For example, by choosing features related to the difficulty of the action, the model can evaluate actions of different difficulties more fairly rather than favoring common actions. Aiming at the curse of dimensionality caused by high dimensions and the failure of ordinary classification algorithms due to imbalance, this paper proposes an adaptive feature selection algorithm based on recursive feature elimination. Through exploratory analysis of the sampling mode and the number of optimal feature spaces on public datasets, it provides a referable value range for subsequent related research.

References

[1] N. Kurpiers and U. G. Kersting, (2017) *"The one-ski-method—effects of an alternative teaching approach on selected movement patterns in alpine skiing"* **Cogent Social Sciences** ***3***(1): 1275958. DOI: 10.1080/23311886.2016.1275958.

[2] T. Li, J. Wang, K. Wiltos, and M. Woźniak, (2024) *"An Intelligent Proofreading for Remote Skiing Actions Based on Variable Shape Basis"* **Mobile Networks and Applications**: 1–14. DOI: 10.1007/s11036-024-02419-4.

[3] J. Lin, (2022) *"Classroom teaching design of alpine skiing based on virtual reality technology"* **Mathematical Problems in Engineering** ***2022***(1): 5721790. DOI: 10.1155/2022/5721790.

[4] Y. Sun, S. Yin, H. Li, L. Teng, and S. Karim, (2019) *"GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster"* **IEEE Access** ***7***: 99254–99262. DOI: 10.1109/ACCESS.2019.2926816.

[5] C. Lin, C.-F. Tsai, and W.-C. Lin, (2023) *"Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study"* **Artificial Intelligence Review** ***56***(2): 845–863. DOI: 10.1007/s10462-022-10186-5.

[6] N. Wang and E. Zhao, (2024) *"A new method for feature selection based on weighted k-nearest neighborhood rough set"* **Expert Systems with Applications** ***238***: 122324. DOI: 10.1016/j.eswa.2023.122324.

[7] J. Yu, L. Zhao, S. Yin, and M. Ivanović, (2024) *"News recommendation model based on encoder graph neural network and bat optimization in online social multimedia art education"* **Computer Science and Information Systems** ***21***(3): 989–1012. DOI: 10.2298/CSIS231225025Y.

[8] J. Liu, Y. Lin, J. Du, H. Zhang, Z. Chen, and J. Zhang, (2023) *"ASFS: A novel streaming feature selection for multi-label data based on neighborhood rough set"* **Applied Intelligence** ***53***(2): 1707–1724. DOI: 10.1007/s10489-022-03366-x.

[9] B. Sang, W. Xu, H. Chen, and T. Li, (2023) *"Active anti-noise fuzzy dominance rough feature selection using adaptive k-nearest neighbors"* **IEEE Transactions on Fuzzy Systems** ***31***(11): 3944–3958. DOI: 10.1109/TFUZZ.2023.3272316.

[10] I. Niño-Adan, I. Landa-Torres, E. Portillo, and D. Manjarres, (2022) *"Influence of statistical feature normalisation methods on K-Nearest Neighbours and K-Means in the context of industry 4.0"* **Engineering Applications of Artificial Intelligence** ***111***: 104807. DOI: 10.1016/j.engappai.2022.104807.

[11] L. Sun, J. Zhang, W. Ding, and J. Xu, (2022) *"Feature reduction for imbalanced data classification using similarity-based feature clustering with adaptive weighted k-nearest neighbors"* **Information Sciences** ***593***: 591–613. DOI: 10.1016/j.ins.2022.02.004.

[12] S. Chormunge and S. Jena, (2018) *"Correlation based feature selection with clustering for high dimensional data"* **Journal of Electrical Systems and Information Technology** *5*(3): 542–549. DOI: 10.1016/j.jesit.2017.06.004.

[13] D. Dhabliya, K. Jain, M. Bargavi, A. Dhablia, J. R. R. Kumar, A. Gupta, S. Pramanik, et al. "Item Selection Using K-Means and Cosine Similarity". In: *AI-Driven Marketing Research and Data Analytics*. IGI Global, 2024, 228–244. DOI: 10.4018/979-8-3693-2165-2.ch013.

[14] D. Thakur and S. Biswas, (2024) *"Permutation importance based modified guided regularized random forest in human activity recognition with smartphone"* **Engineering Applications of Artificial Intelligence** *129*: 107681. DOI: 10.1016/j.engappai.2023.107681.

[15] B. Elkari, Y. Chaibi, T. Kousksou, et al., (2024) *"Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems"* **Energy Reports** *12*: 988–999. DOI: 10.1016/j.egyr.2024.07.002.

[16] Z. Runchi, X. Liguo, and W. Qin, (2023) *"An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects"* **Expert Systems with Applications** *212*: 118732. DOI: 10.1016/j.eswa.2022.118732.

[17] D. Lee and S. Suh, (2025) *"Measuring Income and Wealth Inequality: A Note on the Gini Coefficient for Samples with Negative Values"* **Social Indicators Research** *176*(3): 947–965. DOI: 10.1007/s11205-024-03488-4.

[18] V. Ignatenko, A. Surkov, and S. Koltcov, (2024) *"Random forests with parametric entropy-based information gains for classification and regression problems"* **PeerJ Computer Science** *10*: e1775. DOI: 10.7717/peerj-cs.1775.

[19] S. P. Roy, A. Kasat, et al. "Diabetic prediction with ensemble model and feature selection using information gain method". In: *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE. 2024, 1080–1085. DOI: 10.1109/IDCIoT59759.2024.10467649.

[20] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. M. M. B, P. Jayagopal, and G. T. Dalu, (2023) *"An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database"* **Scientific reports** *13*(1): 13588. DOI: 10.1038/s41598-023-40717-1.

[21] J. Qu and M. J. Zuo, (2010) *"Support vector machine based data processing algorithm for wear degree classification of slurry pump systems"* **Measurement** *43*(6): 781–791. DOI: 10.1016/j.measurement.2010.02.014.

[22] Q. Hu, D. Yu, J. Liu, and C. Wu, (2008) *"Neighborhood rough set based heterogeneous feature subset selection"* **Information sciences** *178*(18): 3577–3594. DOI: 10.1016/j.ins.2008.05.024.

[23] M.-S. Yang and Y. Nataliani, (2017) *"A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy"* **IEEE Transactions on Fuzzy Systems** *26*(2): 817–835. DOI: 10.1109/TFUZZ.2017.2692203.

[24] A. R. Yadav, R. S. Anand, M. Dewal, and S. Gupta. "Analysis and classification of hardwood species based on Coiflet DWT feature extraction and WEKA workbench". In: *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE. 2014, 9–13. DOI: 10.1109/SPIN.2014.6776912.

[25] S. K. Behera and R. Dash, (2022) *"A novel feature selection technique for enhancing performance of unbalanced text classification problem"* **Intelligent Decision Technologies** *16*(1): 51–69. DOI: 10.3233/IDT-210057.

[26] F. Qin, A. M. Zain, K.-Q. Zhou, N. B. Yusup, D. D. Prasetya, R. A. Jalil, Z. Z. Abidin, M. Bahari, Y. Kamin, and M. A. Majid, (2025) *"Hybrid Harmony Search Algorithm Integrating Differential Evolution and Lévy Flight for Engineering Optimization"* **IEEE Access**: DOI: 10.1109/ACCESS.2025.3529714.

[27] H. T. Phuong and N. L. Giang, (2021) *"Fuzzy distance-based filter-wrapper incremental algorithms for attribute reduction when adding or deleting attribute set"* **Vietnam Journal of Science and Technology** *59*(2): 261–274. DOI: 10.15625/2525-2518/59/2/15698.